CS4104 Applied Machine Learning

Regression

Regression versus Classification

- Classification: the output variable takes class labels
- Regression: the output variable takes continuous values

Examples

- Predicting House Value
  - Actual Price: £100,000
  - Predicted 1: £99,950 (Very Good Prediction)
  - Predicted 2: £50,000 (Very Bad Prediction)
- Predicting Car Premium
  - Using Location, Age, History etc

zeshan.khan@nu.edu.pk

Regression Techniques

- Linear Regression
- Non-Linear Regression
- Logistic Regression

zeshan.khan@nu.edu.pk

Linear Regression

Regression

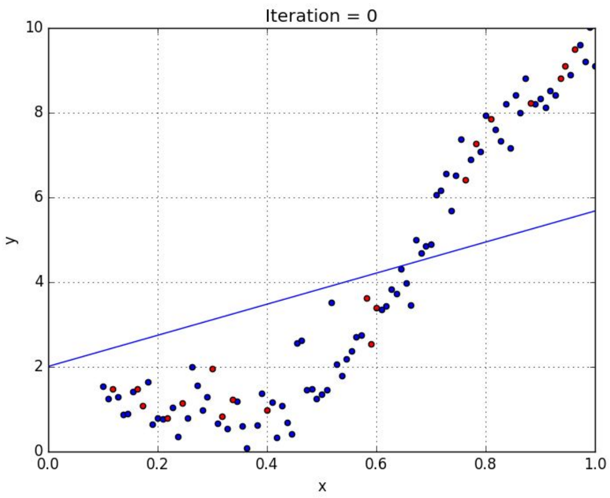
Linear Regression

- Theoretically well motivated algorithm: developed from Statistical Learning Theory

- Empirically good performance: successful applications in many fields (stock prices, insurance etc)

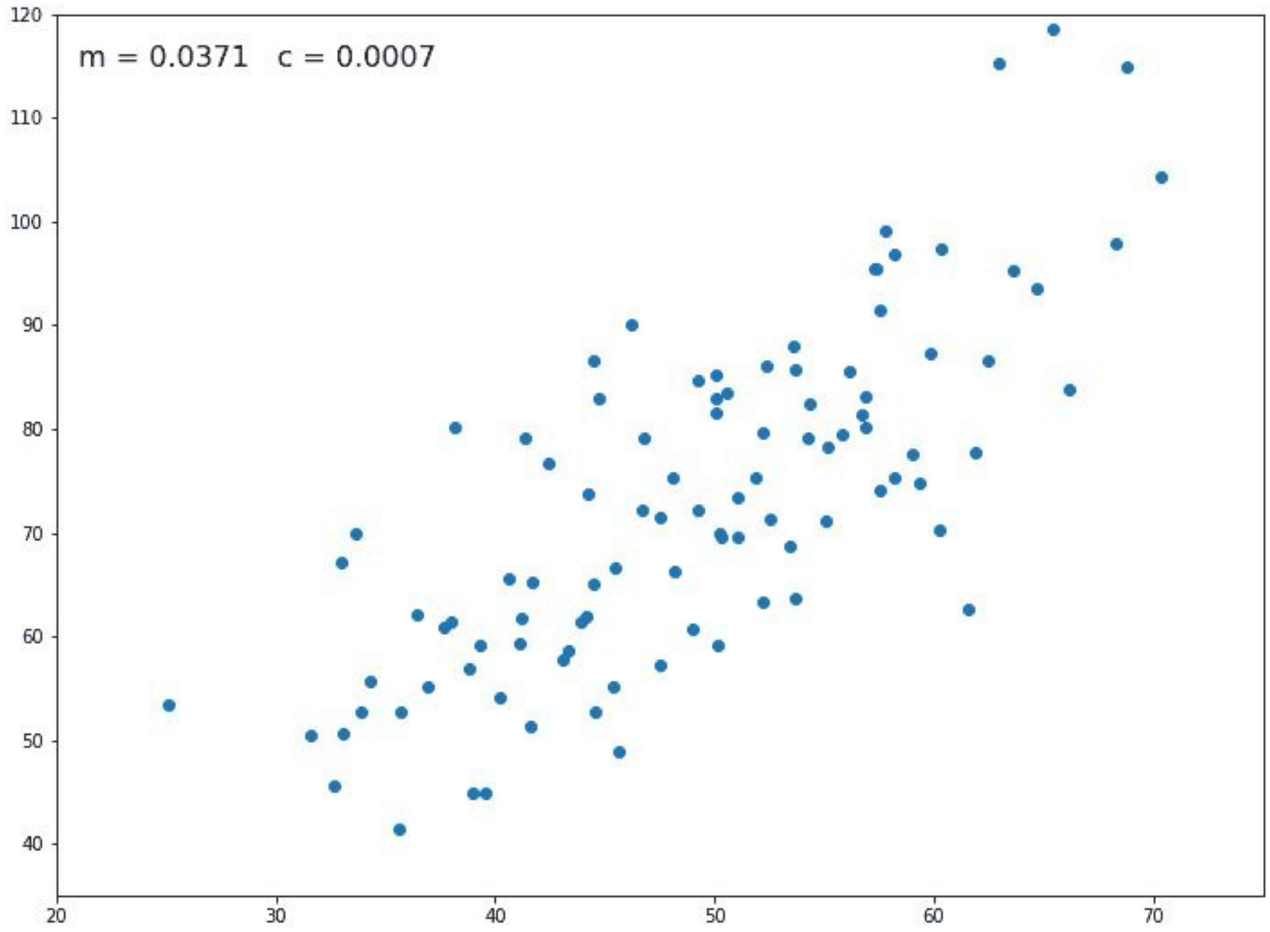
- Given $\{(x_i, y_i) \forall i \in (1, n)\}$

- $predict\ y_k\ for\ x_k \neq x_i\ \forall i \in (1, n)$



zeshan.khan@nu.edu.pk

Regression Line



Formula

- $Y = a + bX$
- $where\ b = r\frac{SD_y}{SD_x}$
- $a = Y' - bX'$
- $slope = \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum X^2 - (\sum X)^2}$

- b is the slope of regression line
- a is the y intercept of regression line
- r is correlation coefficient: a statistic used to describe the strength of the relationship between two variables
- X' and Y' is the mean of X and Y respectively
- SD_x, SD_y are the standard deviation of x and y

zeshan.khan@nu.edu.pk

Example

X	Y
60	3.1
61	3.6
62	3.8
63	4
65	4.1
64	?

zeshan.khan@nu.edu.pk

Step 1: Count

- N=5

zeshan.khan@nu.edu.pk

Step 2: Products

X	Y	X*Y	X*X
60	3.1	186	3600
61	3.6	219.6	3721
62	3.8	235.6	3844
63	4	252	3969
65	4.1	266.5	4225

zeshan.khan@nu.edu.pk

Step 3: Sum

X	Y	X*Y	X*X
60	3.1	186	3600
61	3.6	219.6	3721
62	3.8	235.6	3844
63	4	252	3969
65	4.1	266.5	4225
311	**18.6**	**1159.7**	**19359**

zeshan.khan@nu.edu.pk

Step 4: Slope

- $slope = \frac{N\sum XY-(\sum X)(\sum Y)}{N\sum X^2-(\sum X)^2}$
- $slope = \frac{(5)(1159.7)-(311*18.6)}{(5*19359-311^2)}$
- $slope = 0.1878$

zeshan.khan@nu.edu.pk

Step 5: Intercept

- $Intercept(a) = \frac{\sum Y - b(\sum X)}{N}$
- $Intercept(a) = \frac{18.6 - 0.1878}{5}$
- $Intercept(a) = -7.964$

Step 6: Regression Equation

- $y = a + bx$
- $y = -7.964 + 0.1878x$

zeshan.khan@nu.edu.pk

Step 7: Test Computation

- $y = -7.964 + .01878x$
- $y = -7.964 + .01878 * 64$
- $y = 4.068$

Linear Regression Code

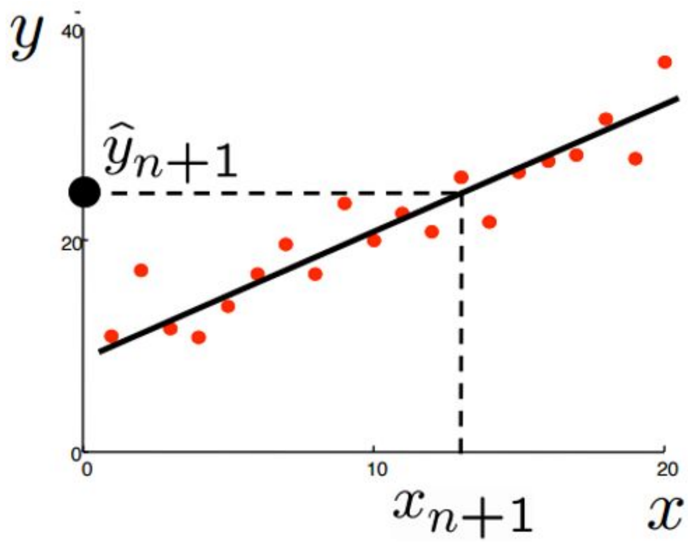
- # importing basic libraries
- import numpy as np, pandas as pd
- from sklearn.model_selection import train_test_split
- train = pd.read_csv('Train.csv')
- test = pd.read_csv('test.csv')
- # importing linear regressionfrom sklearn
- from sklearn.linear_model import LinearRegression
- lreg = LinearRegression()
- splitting into training and cv for cross validation
- x_train, x_cv, y_train, y_cv = train_test_split(X,Y)
- #training the model
- lreg.fit(x_train,y_train)
- #predicting on cv
- pred = lreg.predict(x_cv)
- #calculating mse
- mse = np.mean((pred - y_cv)**2)

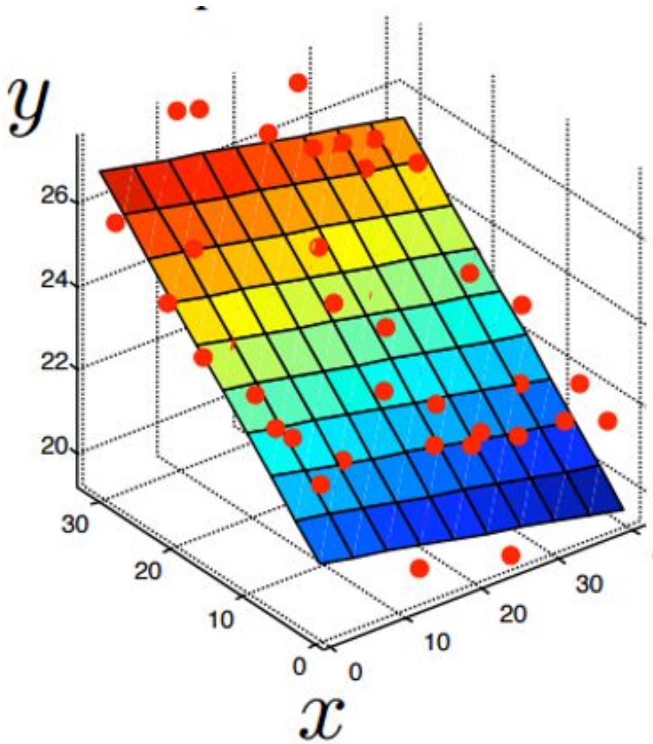
Multivariable Linear Regression

- Estimate y' by a linear function of x:
  - Single variable: $y = a + bx$
  - Multivariable: $y' = a + bx_1 + cx_2 + dx_3 + ex_4$
  - $y' = w_0x_0 + w_1x_1 + w_2x_2 + \cdots + w_dx_d$
  - $y' = w^Tx$
  - w is the parameter to estimate

zeshan.khan@nu.edu.pk

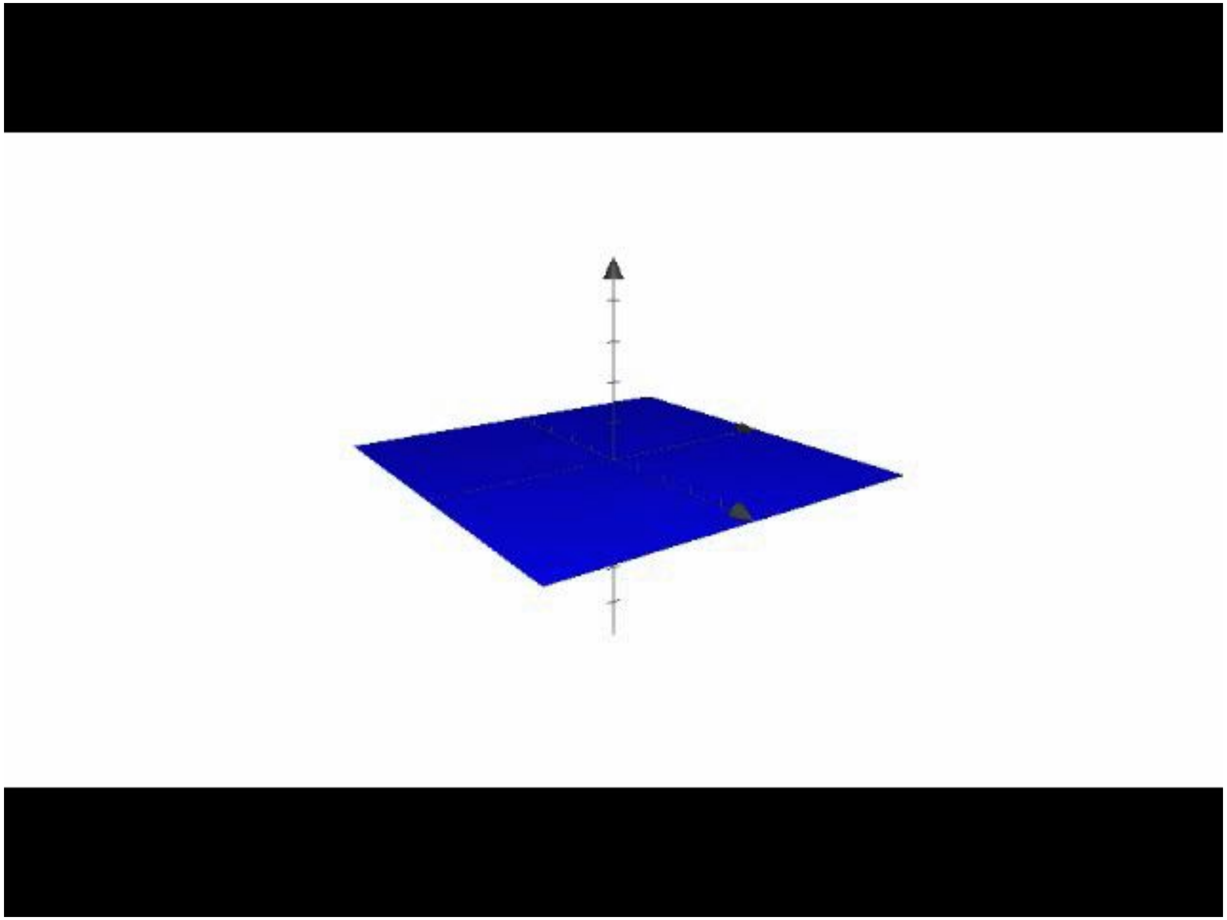
Multivariable Linear Regression





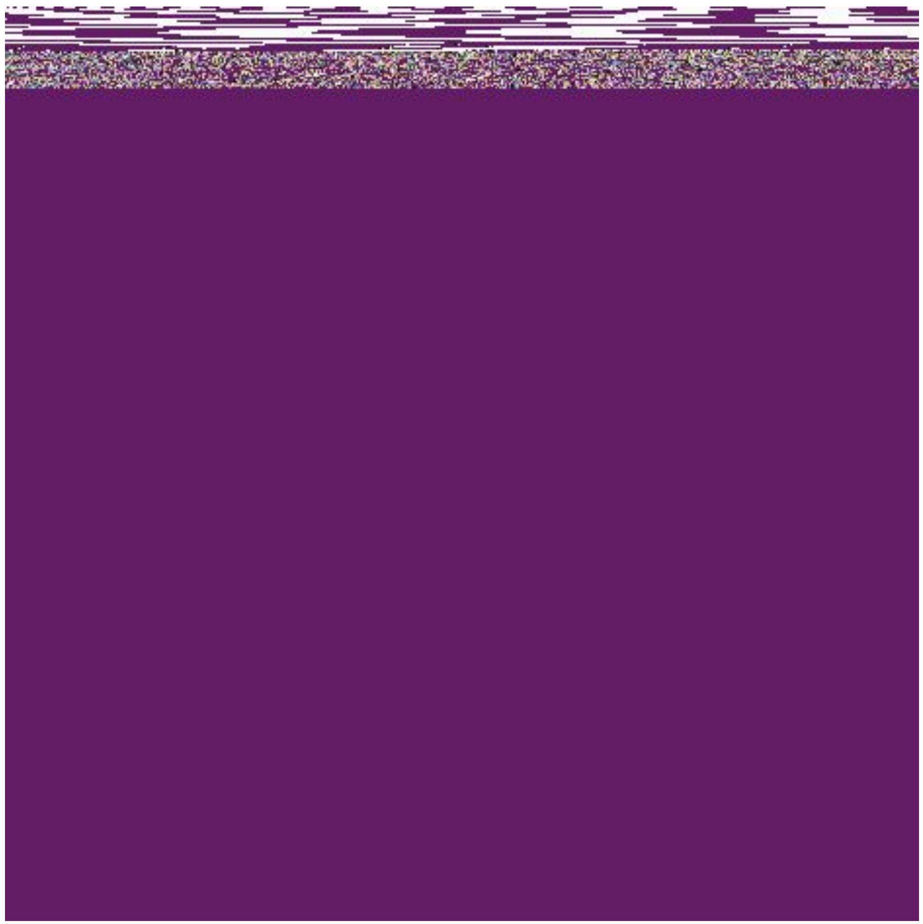
zeshan.khan@nu.edu.pk

Multivariable Linear Regression



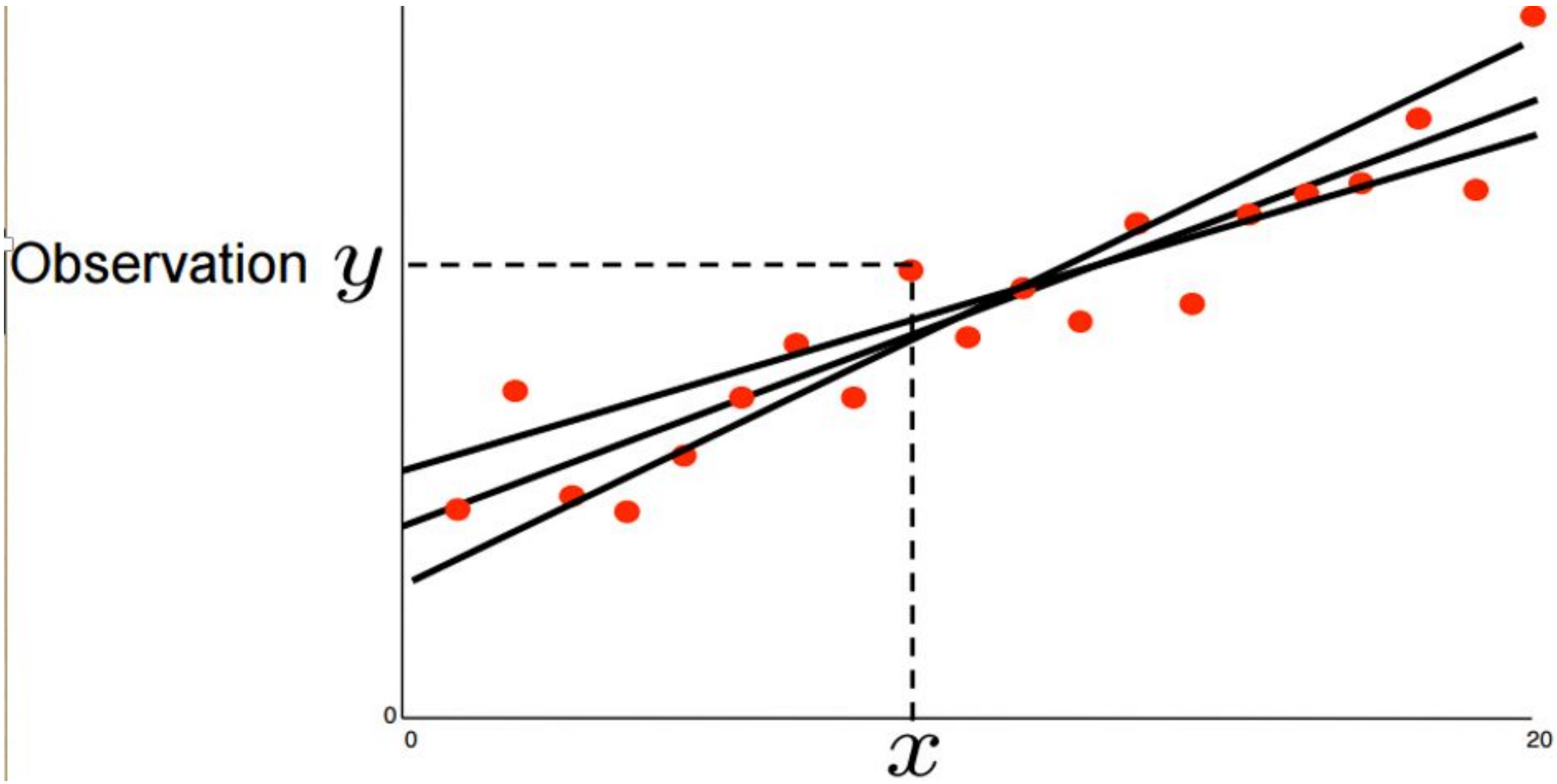
zeshan.khan@nu.edu.pk

Multivariable Linear Regression



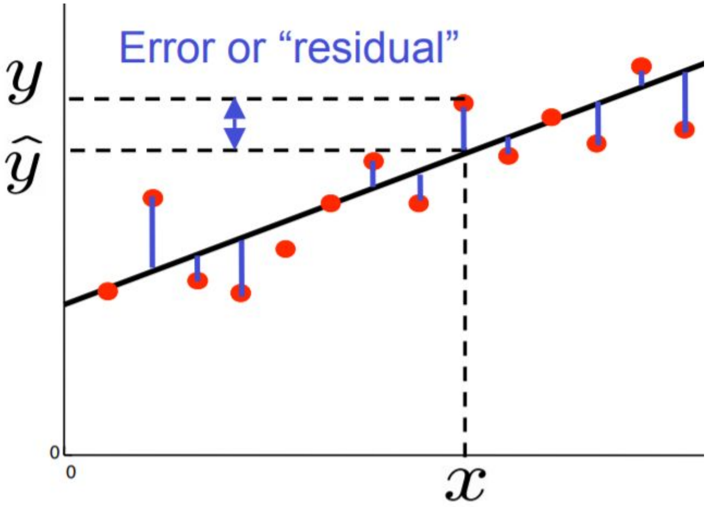
zeshan.khan@nu.edu.pk

Optimal Regressor



Least Mean Square (LMS) Algorithm

- $y' = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_d x_d$
- $predicted_i = y_i{}' = w^T x_i$
- $error_i = E_i = \frac{1}{2}(w^T x_i - y_i)^2$
- $cost = \frac{1}{2}\sum_{i=1}^{n}(w^T x_i - y_i)^2$
- $E = \frac{1}{2}\sum_{i=1}^{n}(w^T x_i - y_i)^2$
- $E = \sum_{i=1}^{n} E_i$
- The objective is to optimize w: E is minimum



zeshan.khan@nu.edu.pk

Evaluation Measure

- Mean Squared Error

Actual (Y)	Predicted (Y')	Y'-Y	Square (Y'-Y)
41	43.6	2.6	6.76
45	44.4	-0.6	0.36
49	45.2	-3.8	14.44
47	46	-1	1
44	46.8	2.8	7.84

Sum of Error = 30.4 /5 = 6.08

zeshan.khan@nu.edu.pk

LMS Algorithm

- By Gradient descent
- $w^{t+1}: w - \alpha \frac{\partial}{\partial w} E$
- $\frac{\partial}{\partial w} E = \sum_{i=1}^{n} \frac{\partial}{\partial w} E_i$
- $\frac{\partial}{\partial w} E_i = \frac{1}{2} (w^T x_i - y_i)^2$
- $\frac{\partial}{\partial w} E_i = (w^T x_i - y_i) * x_i$
- $w_i^{t+1} = w_i^t - \alpha (w^T x_i - y_i) * x_i$

zeshan.khan@nu.edu.pk

Ordinary Least Squares (OLS) Estimate

- $E = \frac{1}{2}\sum_{i=1}^{n}(w^T x_i - y_i)^2$
- In vector form
- $E = \frac{1}{2}(Xw - y)^T(Xw - y)$ Multipliable
- $E = \frac{1}{2}(w^T X^T Xw - 2y^T Xw + y^T y)$ Multipliable
- $\frac{\partial}{\partial w}E = \frac{1}{2} * 2X^T Xw - \frac{1}{2} * 2\ X^T y$
- $\frac{\partial}{\partial w}E = X^T Xw - X^T y$

zeshan.khan@nu.edu.pk

OLS Estimate

- $\frac{\partial}{\partial w} E = X^T X w - X^T y$
- Setting the derivative (change in error) to zero
- $X^T X w - X^T y = 0$
- $X^T X w = X^T y$
- $w = (X^T X)^{-1} X^T y$

Ordinary Least Squares (OLS)

- $w = (X^T X)^{-1} X^T y$
- $if (X^T X)^{-1}$:
  - $unique\ solution$
- $else$:
  - $euclidean\ norm$
  - $ridge\ regression$

OLS Summary

Summary

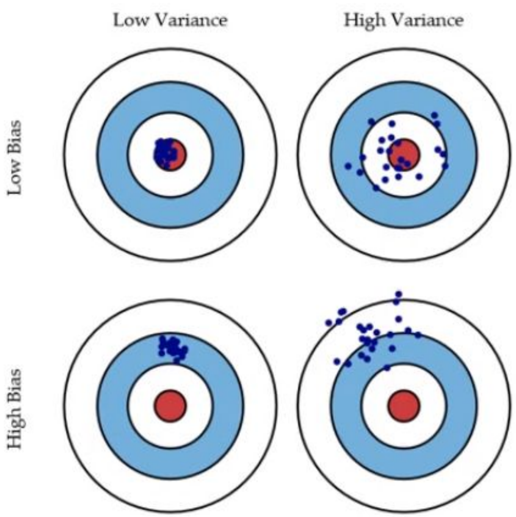
- $solve$:
  - $w = (X^T X)^{-1} X^T y$
- $to\ minimize$:
  - $|Xw - y|^2$

Equivalent Equations

- $E = \frac{1}{2}\sum_{i=1}^{n}(w^T x_i - y_i)^2$
- $E = \frac{1}{2}\sum_{i=1}^{n}(y_i - w^T x_i)^2$
- $E = \frac{1}{2}\sum_{i=1}^{n}(y_i - f(x_i|w))^2$
- $E = \frac{1}{2}\sum_{i=1}^{n}(y_i - f(x_i|w, b))^2$
- $E = \frac{1}{2}\sum_{i=1}^{n}(y_i - x_i\beta_i)^2$

zeshan.khan@nu.edu.pk

Improving the Linear Model



- We may want to improve the simple linear model by replacing Ordinary Least Square (OLS) estimation with some alternative fitting procedure
- Why use an alternative fitting procedure?
  - Prediction Accuracy
  - Model Interpretability
  - Multi-collinearity
  - Overfitting

zeshan.khan@nu.edu.pk

Prediction Accuracy

- If $n \gg p$ the OLS estimates have relatively <u>low bias</u> and <u>low variability</u> especially for n observation when the relationship between the response and predictors (p) is linear

- If $n > p$ then the OLS fit can have high variance and may result in over fitting and poor estimates on unseen observations

- If $n < p$, then the variability of the OLS fit increases dramatically, and the variance of these estimates in infinite

zeshan.khan@nu.edu.pk

Model Interpretability

- When we have a large number of attributes in the model, there will generally be many attributes that have little or no effect on the response
- Including such irrelevant variable leads to unnecessary complexity
- Having these variables in the model makes it harder to see the effect of the important variables
- The model would be easier to interpret by removing (i.e. setting the coefficients to zero) the unimportant variables

zeshan.khan@nu.edu.pk

Multi-collinearity

- Given dataset
  - With features $x_1, x_2, x_3, \ldots, x_d$
- $x_2 \cong function(x_1)$
- Major change in regressor with minor change in data

zeshan.khan@nu.edu.pk

Overfitting

- Carefully selected features can improve model accuracy, but adding too many can lead to overfitting
  - Overfitted models describe random error or noise instead of any underlying relationship
  - They generally have poor predictive performance on test data

zeshan.khan@nu.edu.pk

CS4104 Applied Machine Learning

Polynomial Regression

zeshan.khan@nu.edu.pk

Polynomial Regression

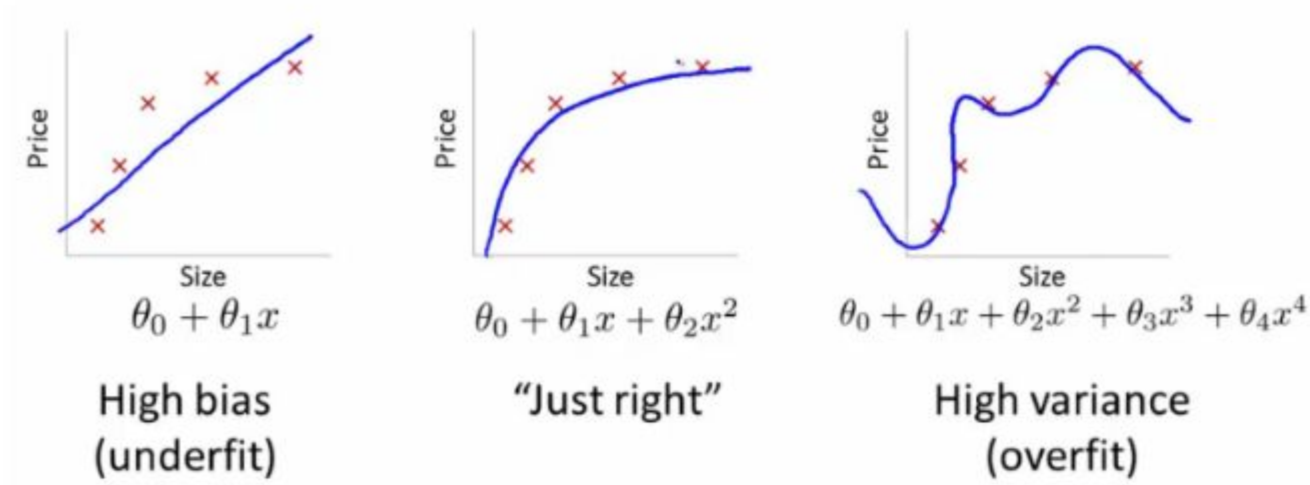
- Linear models become powerful function approximators when we consider non-linear feature transformations.

- $X_i \Rightarrow \begin{bmatrix} a_i \\ b_i \\ c_i \end{bmatrix} \Rightarrow \begin{bmatrix} x_{i0} \\ x_{i1} \\ x_{i2} \end{bmatrix} \Rightarrow \begin{bmatrix} x_i^0 \\ x_i^1 \\ x_i^2 \end{bmatrix}$

- $y_i' = w_0 x_i^0 + w_i x_i^1 + w_2 x_i^2$

zeshan.khan@nu.edu.pk

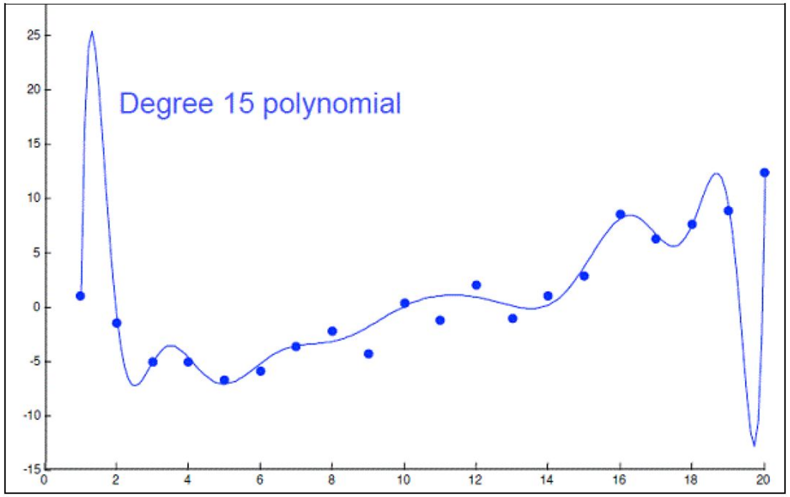
Over-fit: Polynomial Regression



zeshan.khan@nu.edu.pk

Feature/Variable Selection

- For instance, we can use a 15-degree polynomial function to fit the following data so that the fitted curve goes nicely through the data points
- However, a brand new dataset collected from the same population may not fit this particular curve well at all



zeshan.khan@nu.edu.pk

Subset Selection

- Identify a subset of the p predictors that we believe to be related to the response; then, fit a model using OLS on the reduced set.
- Methods: best subset selection, stepwise selection

zeshan.khan@nu.edu.pk

Dimension Reduction

- Involves projecting the p predictors into a M-dimensional subspace, where $M < p$, and fit the linear regression model using the M projections as predictors.
- Methods: principal components regression, partial least squares

zeshan.khan@nu.edu.pk

Shrinkage (Regularization)

- Involves shrinking the estimated coefficients toward zero relative to the OLS estimates
- has the effect of reducing variance and performs variable selection
- Methods: ridge regression, lasso

zeshan.khan@nu.edu.pk

Ridge Regression

Tikhonov Regularization

$$\underset{w}{argmin} \sum_{i} (y_i - w^T x + b)^2$$

Ridge Regression

- $\underset{w,b}{argmin}\, \lambda w^T w + \sum_i (y_i - w^T x + b)^2$
- $regularization = \lambda w^T w$
- $\lambda \gg$ for stable prediction
- $\lambda \gg$ leads to underfit
- Note that $\lambda \geq 0$ is a complexity parameter that controls the amount of shrinkage.
- The idea of penalizing by the sum-of-squares of the parameters is also used in neural networks, where it is known as weight decay.

zeshan.khan@nu.edu.pk

Ridge Regression…

-
- $\underset{\beta}{argmin}\, \lambda\beta^T\beta + \sum_i \left(y_i - \beta_i^T x + \beta_0\right)^2$
- $\underset{\beta}{argmin}\, \lambda\beta^2 + \sum_i \left(y_i - \beta_i^T x + \beta_0\right)^2$
- $\underset{\beta}{argmin}\, \lambda \sum_{j=1} \beta_j^2 + \sum_i \left(y_i - \beta_i^T x + \beta_0\right)^2$

zeshan.khan@nu.edu.pk

Ridge Regression…

- The effect of this equation is to add a shrinkage penalty of the form
- $\lambda \sum_{i=1} \beta_i^2$

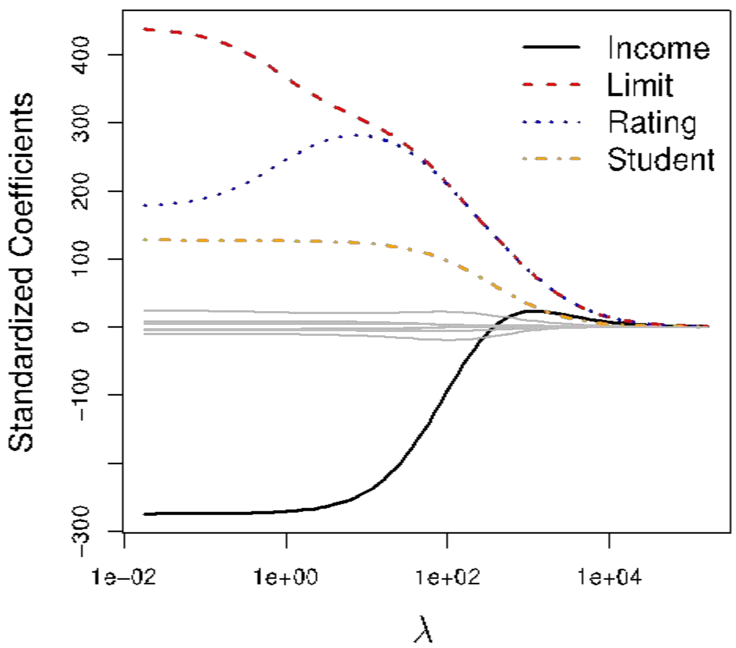
where the tuning parameter λ is a positive value.

- This has the effect of shrinking the estimated beta coefficients towards zero. It turns out that such a constraint should improve the fit, because shrinking the coefficients can significantly reduce their variance
- Note that when $\lambda = 0$, the penalty term as no effect, and ridge regression will procedure the OLS estimates. Thus, selecting a good value for λ is critical (can use cross-validation for this).

zeshan.khan@nu.edu.pk

Ridge Regression…

- As λ increases, the standardized ridge regression coefficients shrinks towards zero.
- Thus, when λ is extremely large, then all of the ridge coefficient estimates are basically zero; this corresponds to the null model that contains no predictors.



zeshan.khan@nu.edu.pk

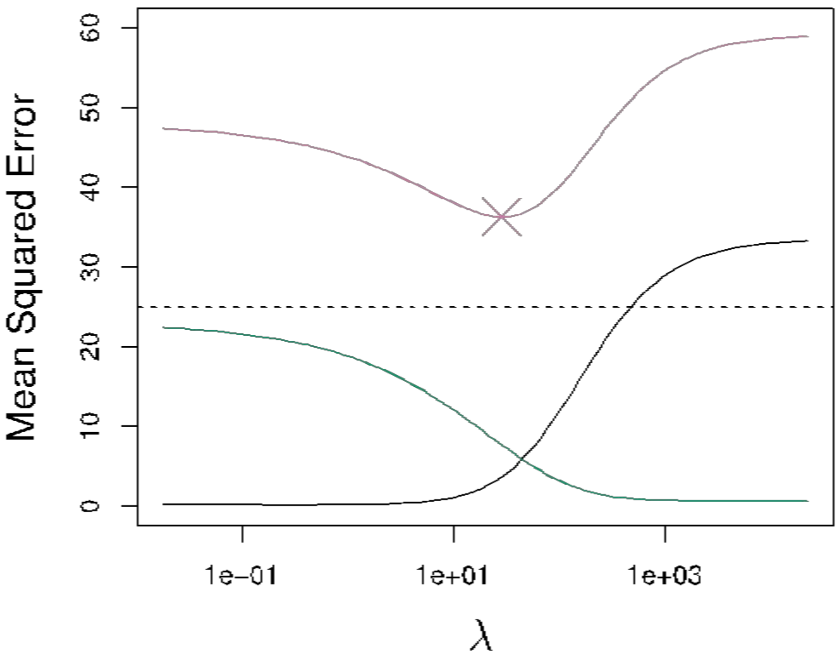
Ridge Regression…

- It turns out that the OLS estimates generally have low bias but can be highly variable. In particular when n and p are of similar size or when $n < p$, then the OLS estimates will be extremely variable
- The penalty term makes the ridge regression estimates *biased* but can also substantially reduce variance
- As a result, there is a bias/variance trade-off.

zeshan.khan@nu.edu.pk

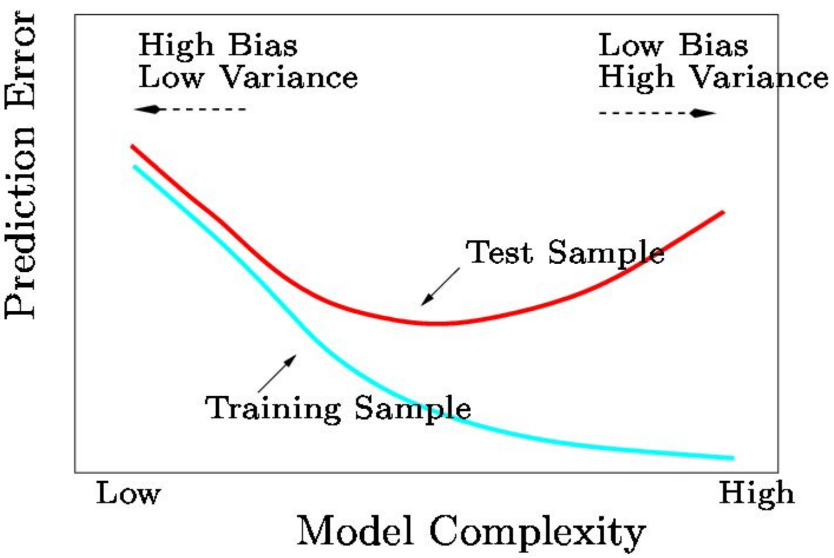
Ridge Regression…

- Black = Bias
- Green = Variance
- Purple = MSE
- Increased λ leads to increased bias but decreased variance



Ridge Regression…

- In general, the ridge regression estimates will be more biased than the OLS ones but have lower variance.
- Ridge regression will work best in situations where the OLS estimates have high variance.



zeshan.khan@nu.edu.pk

Ridge Regression…

- In matrix form:
- *solve*:
  - $\beta = (X^T X + \lambda I)^{-1} X^T y$
- *to minimize*:
  - $|X\beta - y|^2$
- The solution adds a positive constant to the diagonal of $\mathbf{X}^\mathrm{T}\mathbf{X}$ before inversion (making the problem non-singular)

zeshan.khan@nu.edu.pk

Code

- from sklearn.linear_model import Ridge
- #training the model
- ridgeReg = Ridge(alpha=0.05, normalize=True)
- ridgeReg.fit(x_train,y_train)
- pred = ridgeReg.predict(x_cv)
- #calculating mse
- mse = np.mean((pred_cv - y_cv)**2)

zeshan.khan@nu.edu.pk

Lasso

- One significant problem of ridge regression is that the penalty term will never force any of the coefficients to be exactly zero.
- Thus, the final model will include all *p* predictors, which creates a challenge in model interpretation
- A more modern machine learning alternative is the *lasso*.
- The lasso works in a similar way to ridge regression, except it uses a different penalty term that shrinks some of the coefficients exactly to zero.

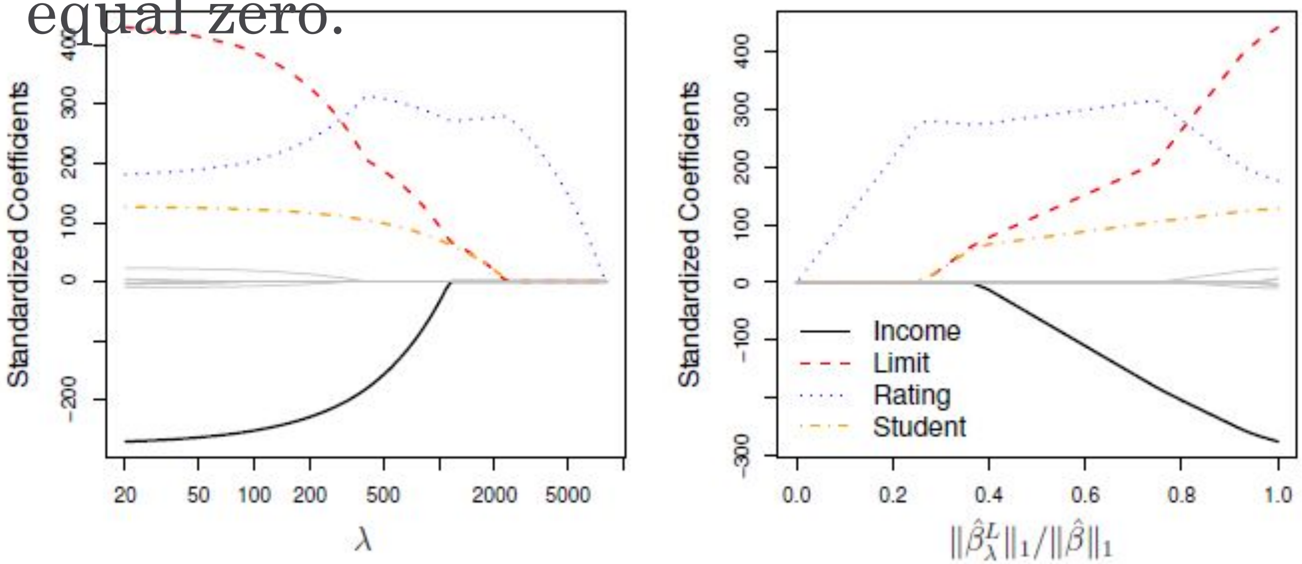
zeshan.khan@nu.edu.pk

Lasso…

- The lasso coefficients minimize the quantity:
- $\underset{\beta}{argmin}\, \lambda \sum_{j=1} |\beta_j \;| + \sum_i \left(y_i - \beta_i^T x + \beta_0\right)^2$
- The key difference from ridge regression is that the lasso uses an ℓ_1 penalty instead of an ℓ_2, which has the effect of forcing some of the coefficients to be exactly equal to zero when the tuning parameter λ is sufficiently large.
- Thus, the lasso performs variable/feature selection.

Lasso…

- When λ = 0, then the lasso simply gives the OLS fit.
- When λ becomes sufficiently large, the lasso gives the null model in which all coefficient estimates equal zero.



zeshan.khan@nu.edu.pk

Lasso vs Ridge

Lasso

- $\lambda \sum_{j=1} |\beta_j \ |$
- Coefficient can be zero/ subset of predictors
- $\sigma \ \propto \frac{1}{\lambda}$
- bias $\propto \lambda$
- High accuracy

Ridge

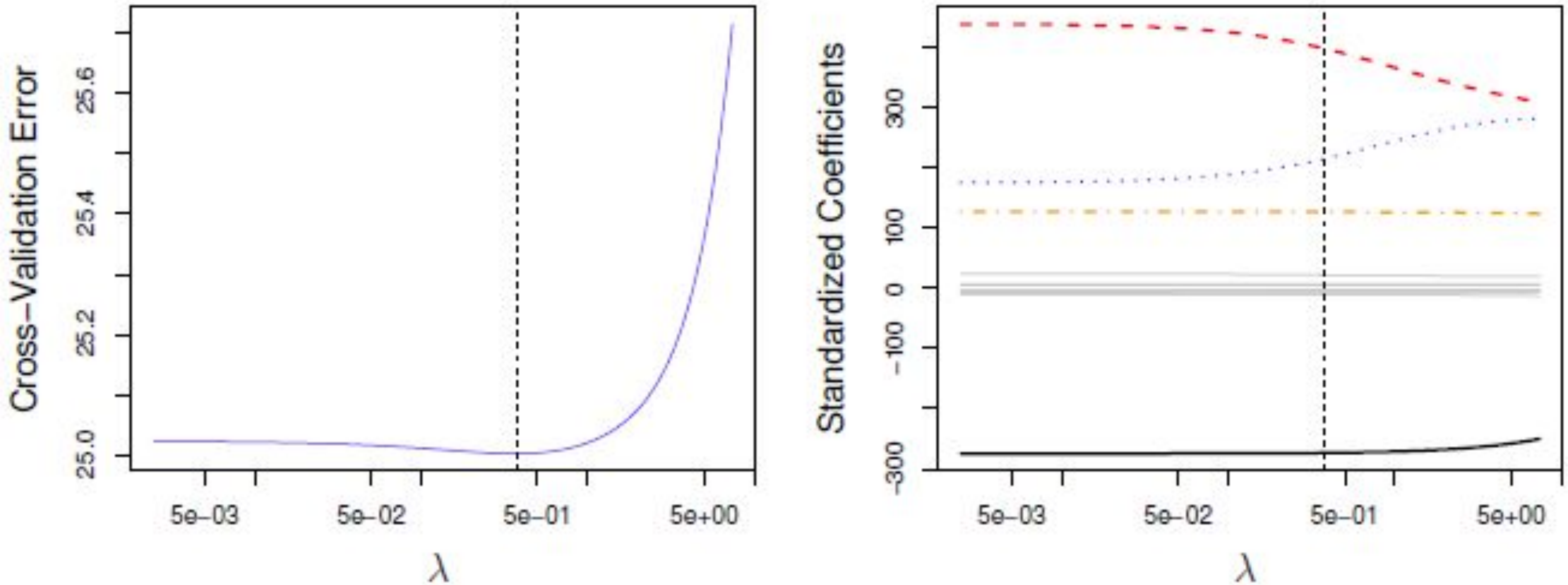
- $\lambda \sum_{j=1} \beta_j^2$
- Coefficient can be smaller but not zero
- $\sigma \ \propto \frac{1}{\lambda}$
- bias $\propto \lambda$
- Relatively low accuracy

zeshan.khan@nu.edu.pk

Selecting the Tuning Parameter λ

- As for subset selection, for ridge regression and lasso we require a method to determine which of the models under consideration in best; thus, we required a method selecting a value for the tuning parameter λ or equivalently, the value of the constraint s.
- Select a grid of potential values; use cross-validation to estimate the error rate on test data (for each value of λ) and select the value that gives the smallest error rate.
- Finally, the model is re-fit using all of the variable observations and the selected value of the tuning parameter λ.

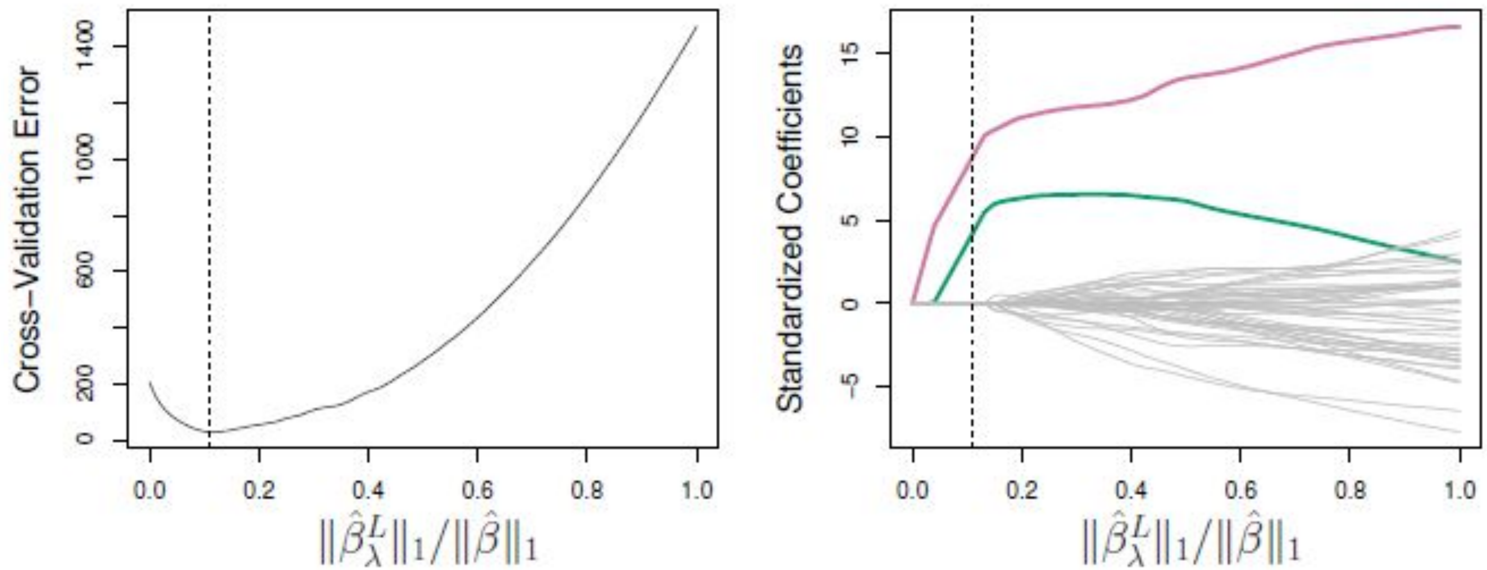
Selecting the Tuning Parameter λ: Credit Data Example



Left: *Cross-validation errors that result from applying ridge regression to the* `Credit` *data set with various values of λ.*
Right: *The coefficient estimates as a function of λ. The vertical dashed lines indicates the value of λ selected by cross-validation.*

zeshan.khan@nu.edu.pk

Selecting the Tuning Parameter λ: Simulated Data Example



Left: *Ten-fold cross-validation MSE for the lasso, applied to the sparse simulated data set from Slide 39.* Right: *The corresponding lasso coefficient estimates are displayed. The vertical dashed lines indicate the lasso fit for which the cross-validation error is smallest.*

Logistic Regression

zeshan.khan@nu.edu.pk

Logistic Regression

- It is an approach for calculating the odds of event happening vs other possibilities…Odds ratio is an important concept
- Goal of logistic regression based classification is to fit the regression curve according to the training data collected (dependent vs independent variables)

zeshan.khan@nu.edu.pk

Why are we studying it?

- To use it for classification
- It is a discriminative classification vs Naïve Bayes' generative classification scheme (what is this?)
- Linear (continuous).. Logistic (categorical): Logit function bridges this gap
- According to Andrew Ng and Michael Jordon logistics regression classification has better error rates in certain situations than Naïve Bayes (eg. large data sets) Big data?

zeshan.khan@nu.edu.pk

Examples

- Mortality of injured patients
- If a patient has a given disease (Recall that we did this using Bayes) (binary classification using a variety of data like age, gender, BMI, blood tests etc.)
- If a person will vote Democratic or Republican
- The odds of a failure of a process, system or a product
- A customer's propensity to purchase a product

zeshan.khan@nu.edu.pk

Examples

- Odds of a person staying in the workforce
- Odds of a homeowner defaulting on a loan
- Conditional Random Field (CRF) an extension of logistic regression to sequential data, are used in NLP. It is labeling a sequence of items so that an entity can be recognized (named entity recognition).
- The appropriate regression analysis to conduct when the dependent variable is dichotomous (binary).
- Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

zeshan.khan@nu.edu.pk

Basics

- odds ratio $= \frac{p}{1-p}$
- Basic function is: logit → logistic regression
- Definition:
- $logit(p) = \log\left(\frac{p}{1-p}\right) = \log(p) - \log(1-p)$
- $\log_b\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_d x_d = \beta_0 + \sum_{i=1}^{d} \beta_i x_i$
- $p = \frac{1}{1+b^{\beta_0+\sum_{i=1}^{d} \beta_i x_i}}$
- The logit function takes x values in the range [0,1] and transforms them to y values along the entire real lines

zeshan.khan@nu.edu.pk