Lecture 3: Linear Regression

MAST30034 Applied Data Science

Dr. Karim Seghouane
School of Mathematics & Statistics
The University of Melbourne

August 8, 2023

Outline

- Linear Regression Model
- Estimation Issues
- Feature Selection
- Significance of regression coefficients
- Model Selection
- Regularization

Linear Regression Model

- Linear regression is a useful tool for predicting a continuous response.
- Response variables: $y_i \in \mathbb{R}$ and p-dimensional vector of predictors (explanatory variables/features), $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top \in \mathbb{R}^p$.
- Given n samples $\{x_i, y_i\}_{i=1}^n$, the aim is to approximate the response variable y_i using a linear combination of the predictors

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i,$$

where $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^\top$, $\mathbb{E}(\epsilon_i) = 0$, and $\mathbb{V}(\epsilon_i) = \sigma^2$.

Goal is to estimate the regression coefficients $\boldsymbol{\beta}$ and σ^2 given the data.

Least square (LS)

How do we fit the linear model to a set of training data?

- We estimate coefficients with $\widehat{\beta}$.
- We choose value of $\widehat{\beta}$ that minimises the residual sum of squares

$$\text{RSS}(\beta) = (Y - X\beta)^T(Y - X\beta),$$

where $Y = (y_1, ..., y_n)^T$, $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ and X is the $n \times (p + 1)$ matrix defined by

$$X = \begin{bmatrix} 1 & x_{11} & \ldots & x_{1p} \\ \vdots & & & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{bmatrix}.$$

Analytical minimization of LS solution

The least square estimator is given by

$$\begin{aligned}\hat{\beta}^{LS} &= \text{argmin}_{\beta}\ \text{RSS}\ (\beta), \\ &= \text{argmin}_{\beta}(Y - X\beta)^T(Y - X\beta), \\ &= (X^TX)^{-1}X^TY \quad \text{only holds if } X \text{ is full rank}\end{aligned}$$

obtained via differentiating w.r.t β and set the first derivative to zero

$$\frac{\partial RSS(\beta)}{\partial \beta} = -2X^T(Y - X\beta).$$

The estimate of the error variance is

$$\widehat{\sigma}^2 = \frac{RSS(\hat{\beta}^{LS})}{n - p - 1}.$$

Assumptions and how do we check them

Recall: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$.

1. The predictors are accurately observed.
2. Linear relationship between the response and predictors.
3. Independent errors and constant variance.
4. X is full rank. This is an important assumption because a full rank X implies that $X^T X$ is invertible and therefore the normal equations $X^T X\beta = X^T Y$ have a unique solution.
   $n > p$ and the $\{x_i\}_{i=1}^n$ are not linear combinations of each other. This can be checked $\det(\mathbb{X}^\top \mathbb{X}) \neq 0$.
   Full rank of $\mathbb{X}$: $n > p$ and the $\{x_i\}_{i=1}^n$ are not linear combinations of each other. Tool: Check that $\det(\mathbb{X}^\top \mathbb{X}) \neq 0$.

Hint: It is VERY important for you to assess the validity of the above assumptions with valid plots and tests.

Least square (LS): Asymptotic

Recall that the least square estimator is given by

$$\hat{\beta}^{LS} = (X^T X)^{-1} X^T Y.$$

- The Gauss-Markov theorem implies $\hat{\beta}^{LS}$ has the smallest mean squared error (MSE) of all linear estimators with NO bias.
- Consistency: $\hat{\beta}^{LS} \to_p \beta$.
- Asymptotic normality. $\hat{\beta}^{LS}$ follows asymptotically a normal distribution by central limit theorem.

Interpreting regression coefficient estimates

Recall: $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + \cdots + \widehat{\beta}_p x_{ip}$.

- Interpretation: 1 unit increase in x_1 corresponds to an estimated $\widehat{\beta}_1$ (unit) increase in the response.
- But the value of $\widehat{\beta}$ changes due to sampling variation.....
- If β_1 is plausibly 0, then we say x_1 is not a significant/useful feature.

Why perform feature selection?

Reasons for feature selection:

- Model is more interpretable to stakeholders. *Imagine asking a doctor to remember the effect of 50 clincial features on MI risk.*
- Higher predictive accuracy. The presence of a large number of non-significant variables may mask effect of significant features.
- Faster computation. Smaller number of features means lower computation time.

How to perform feature selection?

Approaches to feature selection:

- Stepwise regression
- Best subset regression
- Penalised regression

Test significance of regression coefficients

Does the feature X_j predict the response Y?

- We test $H_0 : \beta_j = 0$ (no effect) v.s. $H_a : \beta_j \neq 0$ (there is an effect).
- $T-$test: under H_0, $\hat{\beta}_j / \left(\sqrt{v_j} \hat{\sigma} \right) \sim t_{n-p-1}$
- If $\hat{\beta}_j / \left(\sqrt{v_j} \hat{\sigma} \right) > t_{n-p-1,1-\alpha/2}$ or $\hat{\beta}_j / \left(\sqrt{v_j} \hat{\sigma} \right) < t_{n-p-1,1-\alpha/2}$ we reject H_0
- Unique effect: Usefulness of x_j the other features $x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_p$ are in the model.

Significance of regression coefficients

Recall: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$.

- Unique effect (β_j)
- Marignal effect $y_i = \beta_0 + \beta_j x_{ij} + \epsilon_i$.
- Careful when predictors are dependent
- Independent predictors (PCA)

Significance of regression coefficient

Does x_{ij} affect the response y_i

- $H_0 : \beta_j = 0$ v.s. $H_a : \beta_j \neq 0$
- F-test

$$F = \frac{\left(\frac{RSS_1 - RSS_2}{p_2 - 1}\right)}{\left(\frac{RSS_2}{n - p_2}\right)}$$

  where RSS_1 is the residual sum of squares of intercept model, RSS_2 is our model. p_2 is the number of parameters.
- Under H_0, F follows an F distribution with $(p_2 - 1, n - p_2)$ degrees of freedom

Comparing two nested models

Consider the the pair of nested models:
$\mathcal{M}_0 = \{X_1, \ldots X_q\}$ vs. $\mathcal{M}_1 = \{X_1, \ldots X_q, X_{q+1}, \ldots, X_p\}$.
Are the extra features $X_{q+1}, \ldots, X_p$ necessary?

The F-statistic is

$$F = \frac{(n-p-1)(\text{RSS}_0 - \text{RSS}_1)}{(p-q)\text{RSS}_0} \sim \mathcal{F}_{p-q,n-p-1}.$$

If F is large, we can say that at least one of the features $X_{q+1}, \ldots, X_p$ is significant. Therefore, $\mathcal{M}_1$ is preferred over $\mathcal{M}_0$.

Goodness of fit

Does the model fit the data well?

- R-square (coefficient of determination):

$$R^2 = 1 - \frac{RSS}{SS_{total}},$$

  where $SS_{total} = \sum(y_i - \bar{y})^2$.

- R^2 is between 0 and 1.

- Is larger R^2 better ? Is it enough when comparing different models ?

Other Selection Methods: Selection Criteria

Selection criterion: Akaike information criterion (AIC) and Bayesian information criterion (BIC) are commonly used for variable selection problem.

Negative maximum log-likelihood + a penalty term

$$\text{AIC} = -\log(L(\hat{\beta})) + k, \quad \text{BIC} = -\log(L(\hat{\beta})) + \frac{1}{2}k \log n,$$

where k is model degrees of freedom and n is the number of observations.

The individual magnitude of the AIC/BIC value is not important.

e.g. $\text{AIC}(M_1) = 200$ would not lead to any conclusion but having $\text{AIC}(M_1) = 200$ and $\text{AIC}(M_2) = 150$ suggests M_2 is a better model than M_1.

Selection criteria

Some challenges:

- Simply applying these criterion for an exhaustive search is computationally expensive, even for p is moderately large.
- e.g. $p = 10$ and $k = 7$ which implies $C_7^{10} = 120$ candidate subsets for selection.
- How to find a good model?

Subset Selection

This approach involves identifying a subset of the p predictors that we believe to be related to the response. We then fit a model using least squares on the reduced set of variables.

- To perform **best subset selection**, we fit a separate least squares regression for each possible combination of the p predictors.
- **Forward Stepwise Selection** and **Backward Stepwise Selection** both search through $p(p+1)/2$ models to identify the best one.

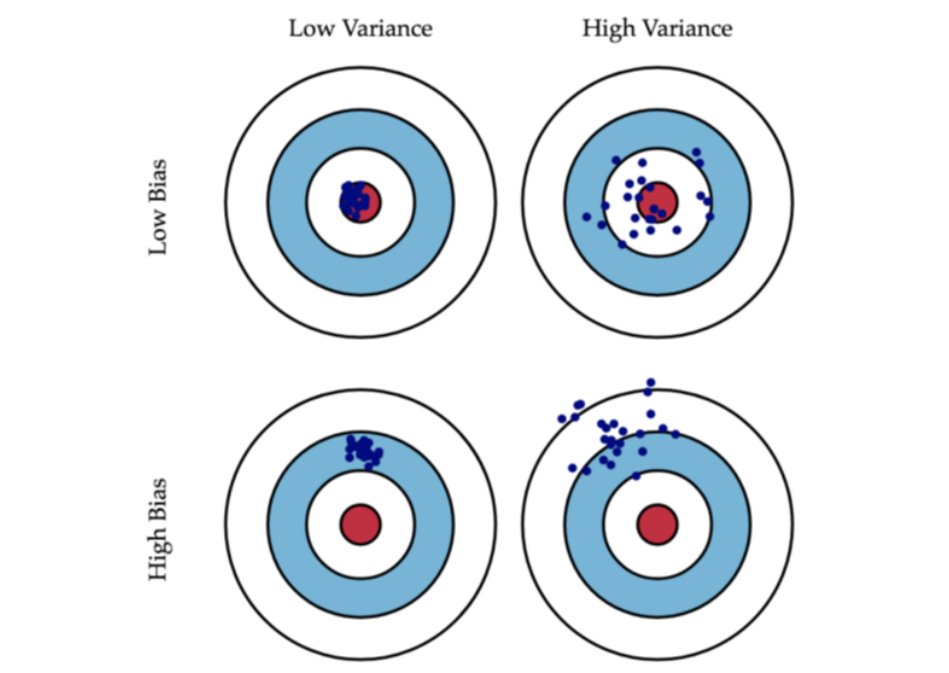
Penalised regression

The idea of shrinkage is to perform a linear regression, while regularizing or shrinking the coefficients $\hat{\beta}$ toward 0.

Why would shrunk coefficients be better ?

- **Bias-variance tradeoff:** Shrinkage introduces bias, but may significantly decrease the variance of the estimates. Modern statistics has explored the trade-off, where it may be worth accepting some bias for a reduction in variance.
- **Result of Shrinking:** Large number of coefficient estimates are zero (Lasso) or close to zero (Ridge) $\rightarrow$ this will help us identify the predictors that exhibit the strongest effects.
- The two best-known techniques for shrinkage are ridge regression and the Lasso.

Bias-variance Illustration



Bias and variance tradeoff of penalised regression estimator

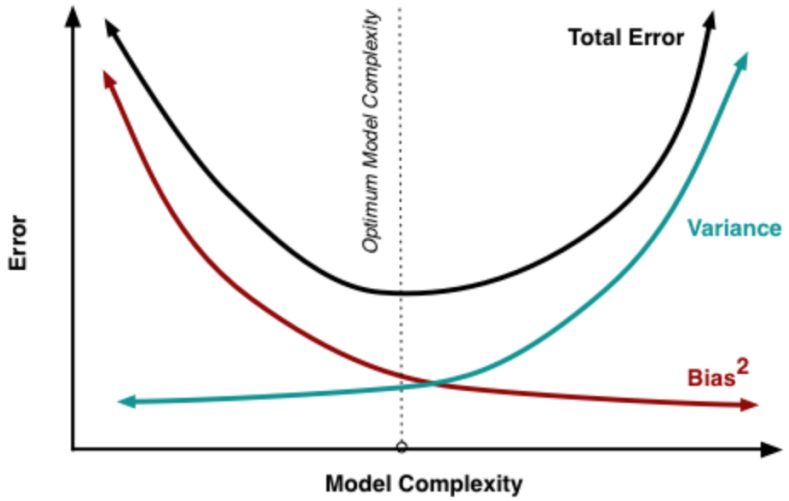
Let $\hat{\beta}$ be an estimator for regression coefficients in the model $Y = X\beta + \epsilon$.

$$\begin{aligned}\text{Total Error}(\hat{\beta}) &= \mathbb{E}\left[Y - X\hat{\beta}\right]^2, \\ &= \left(\mathbb{E}\left[X\hat{\beta}\right] - X\beta)\right)^2 + \mathbb{E}\left[(X\hat{\beta} - \mathbb{E}[X\hat{\beta}])^2\right] + \sigma_\epsilon^2, \\ &= \text{Bias}^2 + \text{ Variance } + \text{ Irreducible Error}.\end{aligned}$$

Bias of $\hat{\beta}^{LS} = 0$. But variance can be very high under multicollinearity or near singularity of $X^\top X$.

Irreducible Error: the noise term in the true relationship that cannot fundamentally be reduced by any model.

Bias-variance trade-off



- Bias is reduced and variance is increased in relation to model complexity.
- Considering overall error, the best spot is the level of complexity at which the increase in bias is equivalent to the reduction in variance. For a model complexity exceeds the best spot, we are over-fitting the model.
- In practice, one need to explore different levels of model complexity and then choose the one minimizing the overall error.

Ridge Regression

Ridge regression solves the following optimization

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

to obtain

$$\hat{\beta}^{Ridge} = \underset{\beta}{\text{argmin}} \left(\text{RSS}(\beta) + \lambda \sum_{j=1}^{p} \beta_j^2 \right)$$

where $\lambda \geq 0$ is a tuning parameter that controls the amount of shrinkage

Solution of Ridge Regression

The *RSS* for ridge regression (after centring inputs X) is expressed as

$$\mathrm{RSS}\left(\beta, \lambda\right) = \left(Y - X\beta\right)^T \left(Y - X\beta\right) + \lambda\beta^T\beta$$

One can obtain a closed-form solution to ridge regression problem with a similar procedure with least square

$$\hat{\beta}^{Ridge} = \left(X^T X + \lambda I_p\right)^{-1} X^T Y.$$

Inclusion of λ makes the problem non-singular even if $X^T X$ is not invertible (singular). This was the original motivation for ridge regression (Hoerl and Kennard, 1970).

Ridge Regression

Ridge regression is like least squares but shrinks the estimated coefficients towards zero.

- As $\lambda \to 0$, $\hat{\beta}^{Ridge} \to \hat{\beta}^{LS}$ and $\lambda \to \infty$, $\hat{\beta}^{Ridge} \to 0$
- Bias of $\hat{\beta}^{Ridge}$ is non-zero! But it is an acceptable tradeoff for lower variance.
- Determining λ is important but also difficult, in practice, where we use cross-validation.
- Note that the intercept term, β_0, is not penalized.
- Standardization: scale each variable before running Ridge is essential, this prevent penalizing some coefficients more than others.

Lasso - Least Absolute Shrinkage and Selection Operator

- The penalty term $\sum_{j=1}^{p} \beta_j^2$ will shrink all of the coefficient towards zero, but it will not set any of them exactly to zero (unless $\lambda = \infty$).
- The final model includes all p predictors.
- This issue is solved by Lasso regression.
- The Lasso coefficients, $\hat{\beta}^{Lasso}$, is obtained by minimizing the quantity

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

Lasso - Least Absolute Shrinkage and Selection Operator

This leads to

$$\hat{\beta}^{Lasso} = \underset{\beta}{\text{argmin}} \left(\text{RSS}(\beta) + \lambda \sum_{j=1}^{p} |\beta_j| \right)$$

where $\lambda \geq 0$ is the tuning parameter.

If $X^\top X = I$, the solution has the form

$$\hat{\beta}_j^{Lasso} = \text{sgn}(\hat{\beta}_j^{LS}) \left(|\hat{\beta}_j^{LS}| - \frac{\lambda}{2} \right)_+ .$$

Shrinkage as constrained optimisation

The method of Lagrange multipliers enables the reformulation of coefficient estimates of ridge and Lasso regression as

$$\hat{\beta}^{Ridge}: \quad \min_\beta \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2$$

$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 \leq t, \qquad \ell_2 \text{ penalty/regularization}$$

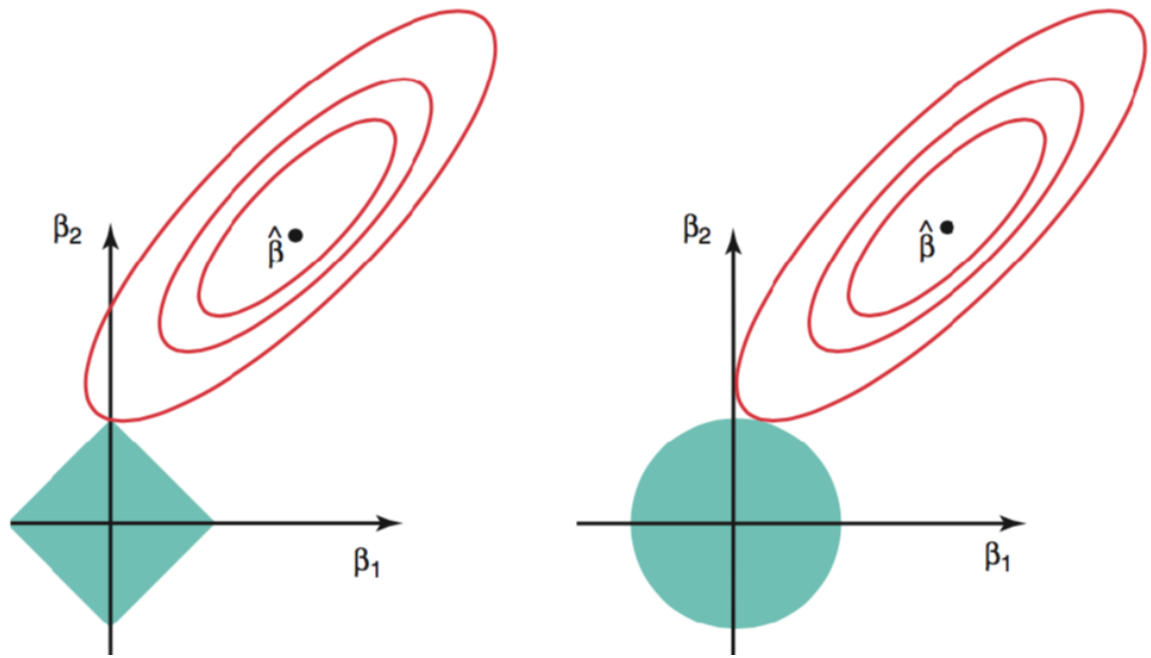
$$\hat{\beta}^{Lasso}: \quad \min_\beta \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2$$

$$\text{subject to } \sum_{j=1}^{p} |\beta_j| \leq t, \qquad \ell_1 \text{ penalty/regularization}$$

for $t \geq 0$; and where there is a one-on-one correspondence between t and tuning parameter λ.

Feasible region and location of solution

Contours of the error and the constraint function for the Lasso and Ridge



t controls the amount of shrinkage. Smaller t implies more shrinkage.

Variable selection property of Lasso

- Why the ℓ_1 shrinkage promotes sparsity?
- Does ℓ_2 shrinkage (ridge regression) also give a sparse solution?

The lasso performs ℓ_1, so that there are "corners" in the constraint, which in two dimensions corresponds to a diamond. If the sum of squares "hits" one of these corners, then the coefficient corresponding to the axis is shrunk to zero.

Ridge regression bring the value of coefficients close to 0 whereas Lasso regression force some of the coefficient to be exactly equal to 0.

Remarks

Regularization in simple terms is a process of introducing additional information in order to solve an ill posed problem or to prevent over-fitting. Elastic net penalty combine both Lasso and Ridge

Pros:

- Lower computational cost than best subset ans stepwise regression.
- Feasible for ultrahigh dimensional problems: $p \gg n$.

Cons:

- Estimates are sensitive to the choice of t or λ.
- Does not do well when true value of coefficients are large and most features are significant.
- Coefficients are biased.