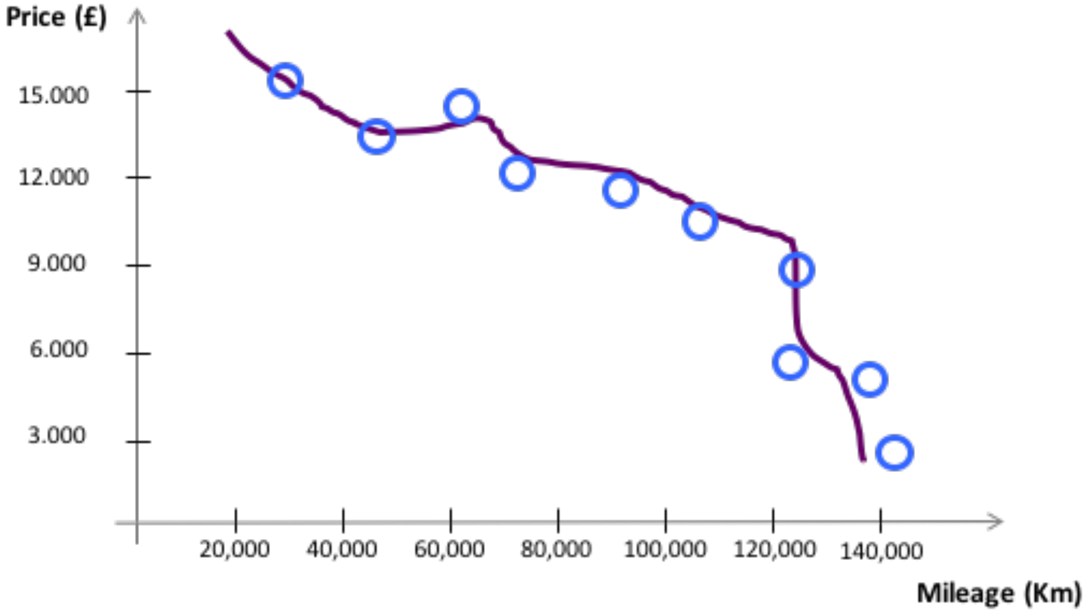
5. Regularization

Regularization

Regularization techniques can address over-fitting by penalizing certain large parameter values

Without Regulatization

$y = \beta_0^* + \beta_1^* x + \beta_2^* x^2 + \beta_3^* x^3 + \beta_4^* x^4$

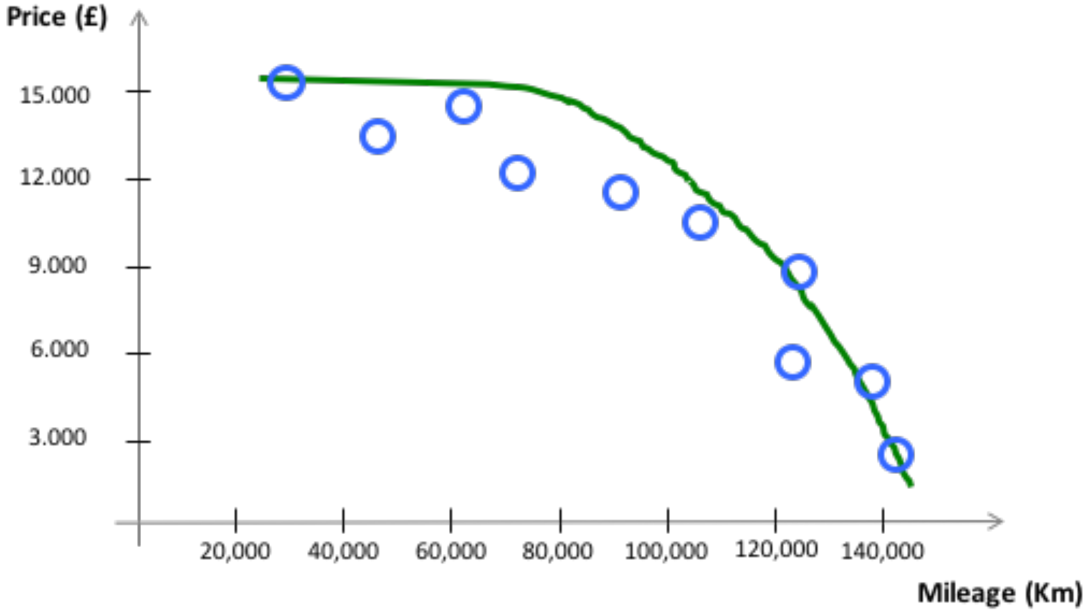


Parameter selection

$$\boldsymbol{\beta}^* = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, tr\big((\boldsymbol{y}-\boldsymbol{X\beta})\cdot(\boldsymbol{y}-\boldsymbol{X\beta})^t\big)$$

With Regularization

$y = \beta_0^* + \beta_1^* x + \beta_2^* x^2 + \beta_3^* x^3 + \beta_4^* x^4$



Parameter selection

$$\boldsymbol{\beta}^* = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, tr\big((\boldsymbol{y}-\boldsymbol{X\beta})\cdot(\boldsymbol{y}-\boldsymbol{X\beta})^t\big) + 10^3\beta_3^2 + 10^3\beta_4^2$$

Regularization

We do not want manual tuning of the importance of the model parameters

Without Regularization

$$(\alpha^*, \beta^*) = \underset{\alpha,\beta}{\operatorname{argmin}}\, SSR(\alpha, \beta) = \underset{\alpha,\beta}{\operatorname{argmin}} \frac{1}{n} \cdot \sum_{i=1}^{n} \left(y_i - (\alpha + \beta x_i)\right)^2$$

With Regularization

$$(\alpha^*, \beta^*) = \underset{\alpha,\beta}{\operatorname{argmin}}\, SSR(\alpha, \beta) = \underset{\alpha,\beta}{\operatorname{argmin}} \frac{1}{n} \cdot \left[\sum_{i=1}^{n} \left(y_i - (\alpha + \beta x_i)\right)^2 + \boxed{\lambda} \cdot \beta^2\right]$$

Regularization parameter

Regularization: Ridge Regression

Simple Linear Regression

Parameter estimates

$$\boldsymbol{\beta}^* = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, tr\big((\boldsymbol{y} - \boldsymbol{X\beta}) \cdot (\boldsymbol{y} - \boldsymbol{X\beta})^t\big)$$

$$\boldsymbol{\beta}^* = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y}$$

Predictions

$$\widehat{\boldsymbol{y}} = \boldsymbol{x} \cdot (\boldsymbol{X}^t\boldsymbol{X})^{-1}\, \boldsymbol{X}^t\boldsymbol{y}$$

Ridge Regression

Parameter estimates

$$\boldsymbol{\beta}^* = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, tr\big((\boldsymbol{y} - \boldsymbol{X\beta}) \cdot (\boldsymbol{y} - \boldsymbol{X\beta})^t\big) + \lambda \cdot \boldsymbol{\beta}^t\boldsymbol{\beta}$$

$$\boldsymbol{\beta}^* = (\boldsymbol{X}^t\boldsymbol{X} + \lambda \cdot \boldsymbol{I})^{-1}\boldsymbol{X}^t\boldsymbol{y}$$

($\boldsymbol{I}$ is the identity matrix)

Predictions

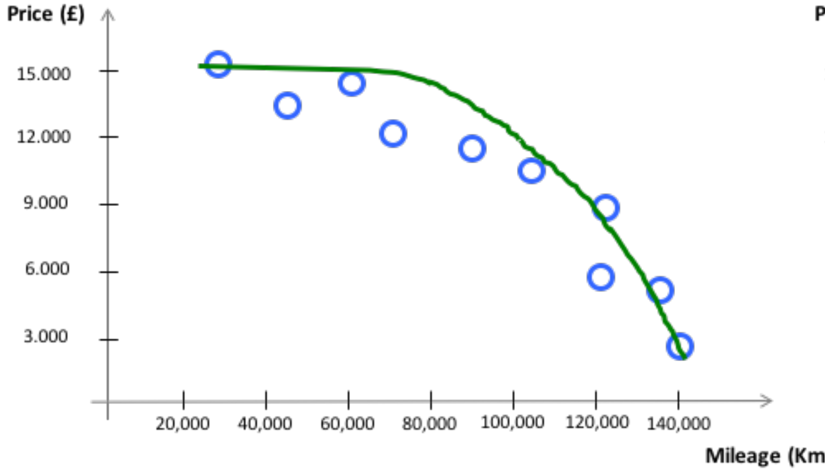
$$\widehat{\boldsymbol{y}} = \boldsymbol{x} \cdot (\boldsymbol{X}^t\boldsymbol{X} + \lambda \cdot \boldsymbol{I})^{-1}\boldsymbol{X}^t\boldsymbol{y}$$

NB: the matrix $\boldsymbol{X}^t\boldsymbol{X} + \lambda \cdot \boldsymbol{I}$ is always invertible whereas the matrix $\boldsymbol{X}^t\boldsymbol{X}$ may not be invertible (e.g. if the number of features is higher than the number of examples / samples.

Regularization: Ridge Regression

What is the role of the regularization parameter λ?

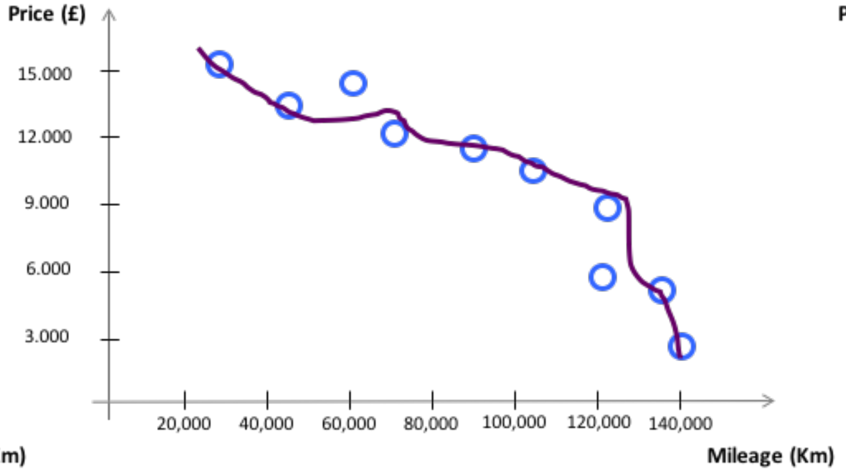
Reasonable λ



Parameter estimates

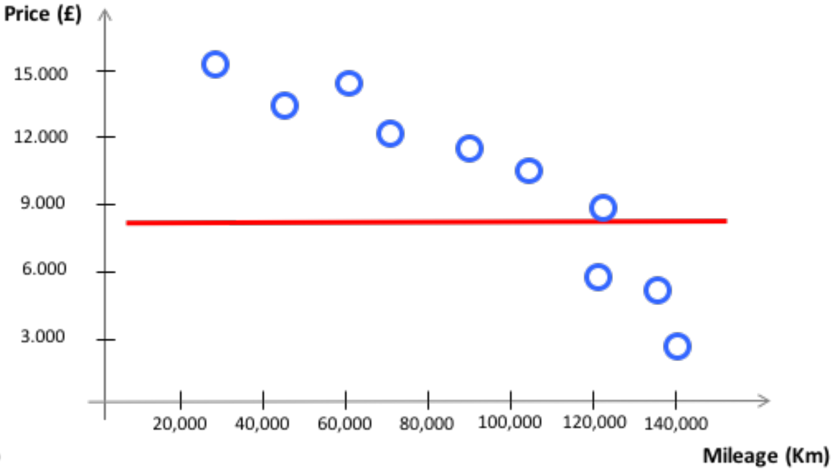
$$\boldsymbol{\beta}^* = (\boldsymbol{X}^t\boldsymbol{X} + \lambda \cdot \boldsymbol{I})^{-1}\boldsymbol{X}^t\boldsymbol{y}$$

Very small λ



Very small $\lambda \Rightarrow$ the optimal parameters for ridge regression correspond to those of simple linear regression

Very large λ



Very large $\lambda \Rightarrow$ the optimal parameters for ridge regression become approximately equal to zero

We need a disciplined way to select the regularization parameter $\Longrightarrow$ **Cross-Validation**

Regularization: LASSO – Least Absolute Shrinkage and Selection Operator

Ridge Regression

Parameter estimates

$$\boldsymbol{\beta}^* = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, tr\big((\boldsymbol{y} - \boldsymbol{X\beta}) \cdot (\boldsymbol{y} - \boldsymbol{X\beta})^t\big) + \lambda \cdot \|\boldsymbol{\beta}\|_2^2$$

$$\Downarrow$$

$$\boldsymbol{\beta}^* = (\boldsymbol{X}^t\boldsymbol{X} + \lambda \cdot \boldsymbol{I})^{-1}\boldsymbol{X}^t\boldsymbol{y}$$

Predictions

$$\hat{\boldsymbol{y}} = \boldsymbol{x} \cdot (\boldsymbol{X}^t\boldsymbol{X} + \lambda \cdot \boldsymbol{I})^{-1}\boldsymbol{X}^t\boldsymbol{y}$$

LASSO

Parameter estimates

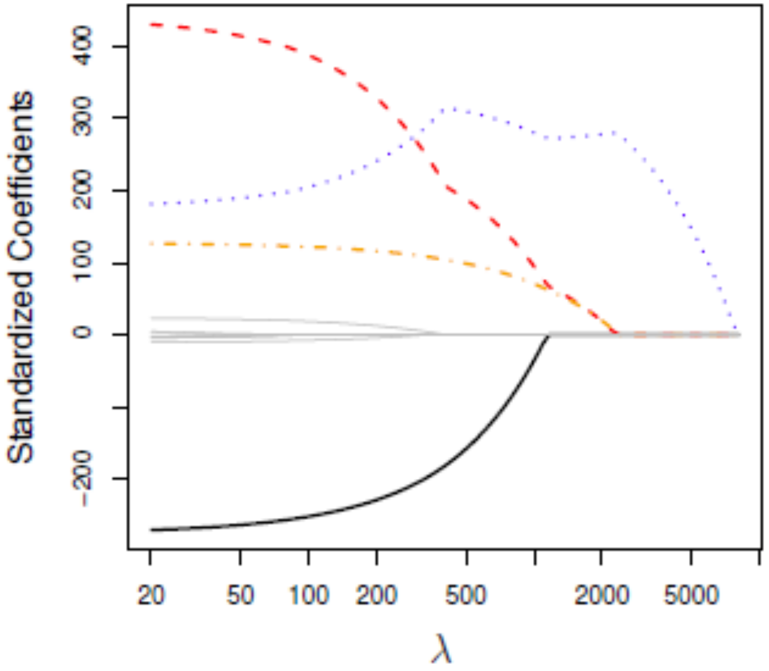
$$\boldsymbol{\beta}^* = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, tr\big((\boldsymbol{y} - \boldsymbol{X\beta}) \cdot (\boldsymbol{y} - \boldsymbol{X\beta})^t\big) + \lambda \cdot \|\boldsymbol{\beta}\|_1$$

The key difference between LASSO and ridge regression is that the LASSO uses an L1 instead of an L2 penalty

This forces some of the coefficients to be set to be equal to zero provided that the hyper-parameter λ is sufficiently large.

Thus, the lasso performs variable/feature selection and shrinkage at the same time.

Regularization: LASSO – Least Absolute Shrinkage and Selection Operator



When λ = 0 , the LASSO solution reduces to the simple linear regression solution.

When λ → ∞ , the LASSO solution is such that the coefficient estimates approach zero.

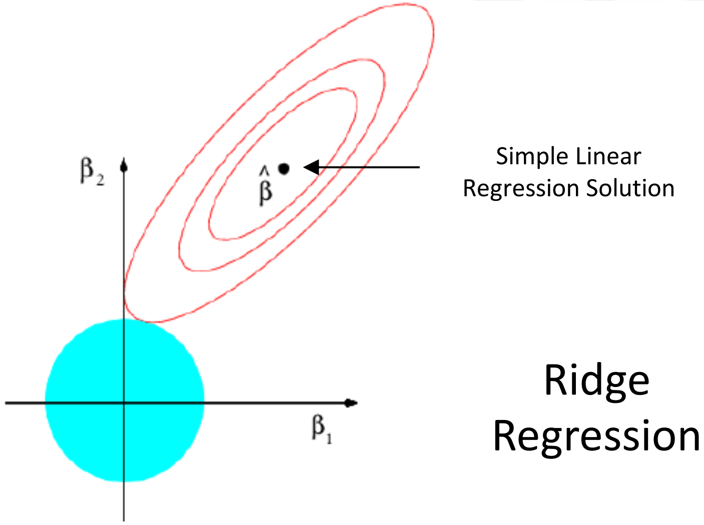
LASSO vs Ridge Regression

Ridge Regression

The ridge regression problem can be re-expressed as follows:

$$\boldsymbol{\beta}^* = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, tr\big((\boldsymbol{y} - \boldsymbol{X\beta}) \cdot (\boldsymbol{y} - \boldsymbol{X\beta})^t\big)\ s.t.\ \|\boldsymbol{\beta}\|_2^2 \leq s$$

where the value of s depends on the value of the hyper-parameter λ.



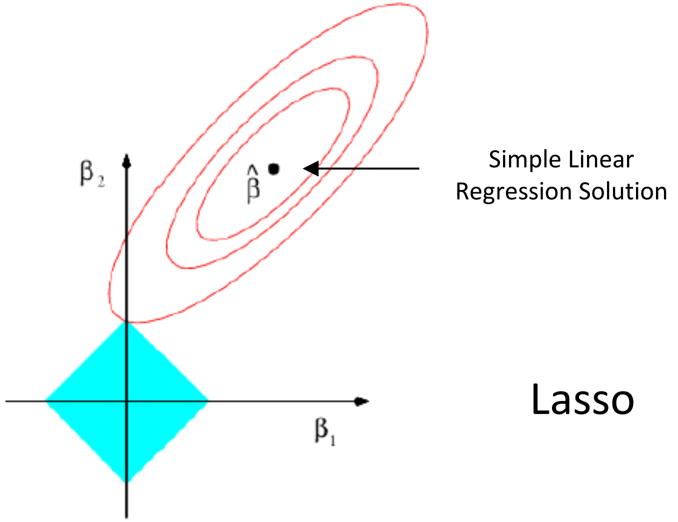
Ridge Regression

LASSO

The ridge regression problem can be re-expressed as follows:

$$\boldsymbol{\beta}^* = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, tr\big((\boldsymbol{y} - \boldsymbol{X\beta}) \cdot (\boldsymbol{y} - \boldsymbol{X\beta})^t\big)\ s.t.\ \|\boldsymbol{\beta}\|_1 \leq s$$

where the value of s depends on the value of the hyper-parameter λ.



Lasso

The LASSO and ridge regression coefficient estimates are given by the first point at which the ellipse contacts the constraint region.

LASSO vs Ridge Regression

- LASSO has an advantage over ridge regression, in that it produces simpler and more interpretable models that involve only a subset of predictors.

- LASSO leads to qualitatively similar behavior to ridge regression, in that as λ increases, the variance decreases and the bias increases.

- LASSO can generate more accurate predictions compared to ridge regression.

- Cross-validation can be used in order to determine which approach is better on a particular data set.