8.2.3 Shrinkage Methods

Multivariate linear regression has low bias, but high variance. Shrinkagemethods try to minimize the overall error by increasing the bias slightly,

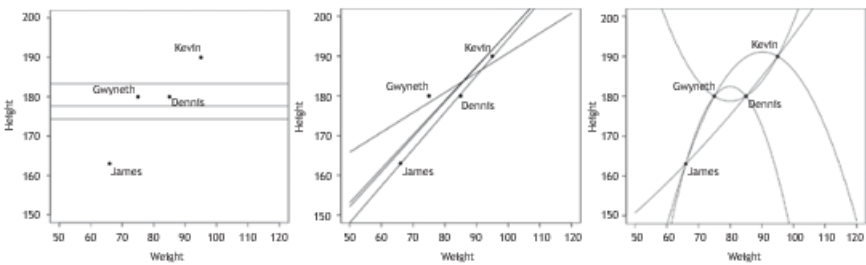


Figure 8.9 Models from the Figure ??: left, average models; center, linear regression models; right, polynomial models.

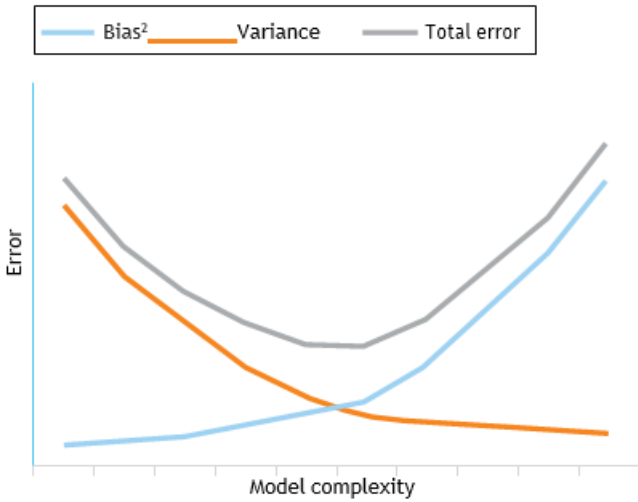


Figure 8.10 The bias-variance trade-off.

while reducing the variance component of the error. Two of the best-known shrinkage methods are ridge and lasso regression.

8.2.3.1 Ridge Regression

Ridge regression increases the bias component of the overall error by adding a penalty term for the coefficients $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ to Equation (8.10), leading to the following objective function for optimization:

$$\underset{\hat{\beta}_0,\ldots,\hat{\beta}_p}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} error(y_i, \hat{y}) + \lambda \times \sum_{j=1}^{p} \hat{\beta}_j^2 \right\} \tag{8.13}$$

for n instances with p predictive attributes. This can be, according to Equation (8.8), rewritten as:

$$\underset{\beta_0,\ldots,\beta_p}{\operatorname{argmin}} \left(\sum_{i=1}^{n} \left[y - \hat{\beta} + \sum_{j=1}^{p} \beta_j \times x_{ij} \right]^2 + \lambda \times \sum_{j=1}^{p} \hat{\beta}_j^2 \right)$$

Assessing and evaluating results As for MLR, the main result of ridge regression is a set of estimates for the $\hat{\beta}_j$ coefficients. Indeed, ridge regression is also a multivariate linear model, but uses a different method to learn the $\hat{\beta}_j$ coefficients.

Setting the hyper-parameters Ridge regression has one hyper-parameter, the λ,that penalizes the $\hat{\beta}_j$ coefficients, i.e., as larger λ is, more costly is to have larger $\hat{\beta}_j$ coefficients. The right value for λ is problem dependent.

Table 8.3 Advantages and disadvantages of ridge regression.

Advantages	Disadvantages
• Strong mathematical foundation • Easily interpretable • Deals better with correlated predictive attributes than ordinary least squares.	• Number of instances must be larger than number of attributes • Sensitive to outliers • Data should be normalized • When relation between predictive and the target attributes is non-linear, uses information poorly

Advantages and disadvantages of ridge regression The advantages and disadvantages of ridge regression are shown in Table 8.3.

8.2.3.2 Lasso Regression

The least absolute shrinkage and selection operator (lasso) regression algorithm is another penalized regression algorithm, that can deal efficiently with high-dimensional data sets. It performs attribute selection by taking into account not only the predictive performance of the induced model, but also the complexity of the model. The complexity is measured by the number of predictive attributes used by the model. It does this by including in the equation of the multivariate linear regression model an additional weighting term, which depends on the sum of the $\hat{\beta}_j$ weights modules. The weight values define the importance and number of predictive attributes in the induced model.

The lasso algorithm usually produces sparse solutions. Sparse means that a large number of predictive attributes have zero weight, resulting in a regression model that uses a small number of predictive attributes. As well as attribute selection, the lasso algorithm also performs shrinkage. Mathematically, the lasso algorithm is very well founded.

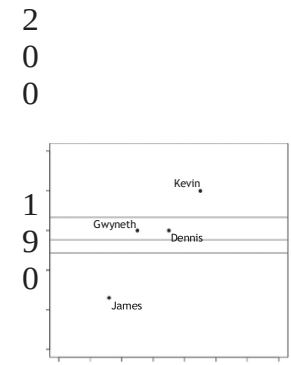
The lasso formulation is quite similar to the ridge formulation, as presented in Equations (8.14) and (8.13):

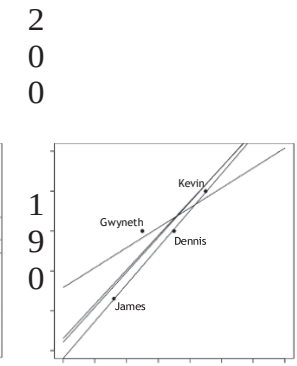
$$\underset{\hat{\beta}_0,\ldots,\hat{\beta}_p}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} error(y_i, \hat{y}) + \lambda \times \sum_{j=1}^{p} |\hat{\beta}_j| \right\} \tag{8.15}$$

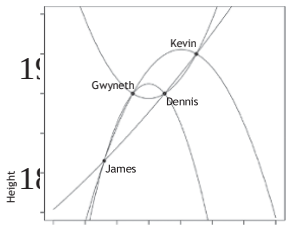
However, they result in substantially different models. This is because the lasso formulation favors the existence of many zero $\hat{\beta}_j$ coefficients. To illustrate why this happens, let us assume that $\hat{\beta}_1 = 0.2$ and $\hat{\beta}_2 = 0.3$. The ridge approach will have $\sum_{j=1}^{p} \hat{\beta}_j^2 = 0.2^2 + 0.3^2 = 0.04 + 0.09 = 0.13$ while the lasso approach will have $\sum_{j=1}^{p} |\hat{\beta}_j| = |0.2| + |0.3| = 0.5$. But if, instead, $\hat{\beta}_1 = 0.5$ and $\hat{\beta}_2 = 0$, ridge regression will have $\sum_{j=1}^{p} \hat{\beta}_j^2 = 0.5^2 + 0^2 = 0.25 + 0 = 0.25$, a value larger than

Table 8.4 Advantages and disadvantages of the lasso.

Advantages	Disadvantages
• Strong mathematical foundation • Easier interpretation than ordinary least squares or ridge regression because it produces simpler models (with fewer predictive attributes) • Deals better with correlated predictive attributes than ridge regression or ordinary least squares; • Automatically discounts irrelevant attributes	• Number of instances must be larger than number of attributes • Sensitive to outliers • Data should be normalized • When the relationship between predictive and target attributes is non-linear, uses information poorly


Kevin
Gwyneth
Dennis
James


Kevin
Gwyneth
Dennis
James


Kevin
Gwyneth
Dennis
James

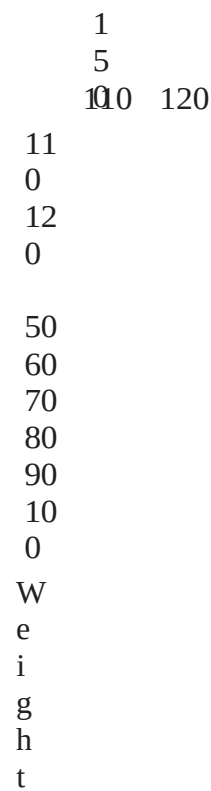


Figure 8.9 Models from the Figure **??**: left, average models; center, linear regression models; right, polynomial models.

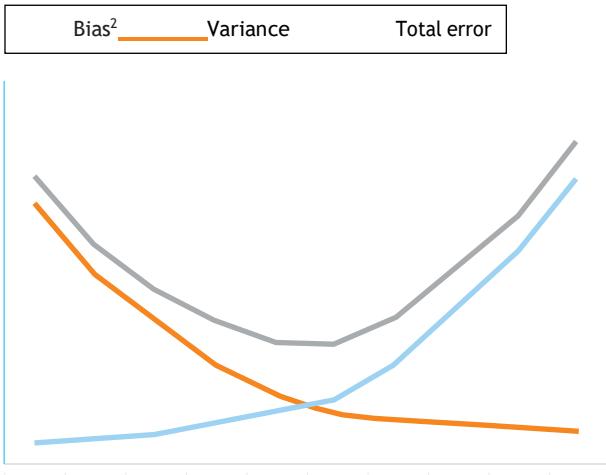


Figure 8.10 The bias–variance trade-off.

while reducing the variance component of the error. Two of the best-known shrinkage methods are ridge and lasso regression.

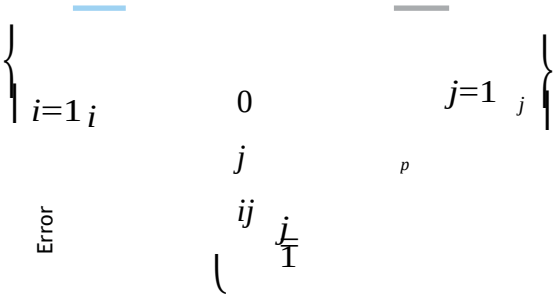
8.2.3.1 Ridge Regression

Ridge regression increases the bias component of the overall error by adding apenalty term for the coefficients $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ to Equation (8.10), leading to the following objective function for optimization:

$$\underset{\hat{\beta}_0,\ldots,\hat{\beta}_p}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} error(y_i, \hat{y}_i) + \lambda \times \sum_{j=1}^{p} \hat{\beta}_j^2 \right\} \tag{8.13}$$

for n instances with p predictive attributes. This can be, according to Equation (8.8), rewritten as:

$$\underset{\hat{\beta}_0,\ldots,\hat{\beta}_p}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left[y_i - \left(\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j \times x_{ij} \right) \right]^2 + \lambda \times \sum_{j=1}^{p} \hat{\beta}_j^2 \right\} \tag{8.14}$$



Assessing and evaluating results As for MLR, the main result of ridge regression is a set of estimates for the $\hat{\beta}_j$ coefficients. Indeed, ridge regression is also a mul- tivariate linear model, but uses a different method to learn the $\hat{\beta}_j$ coefficients.

Setting the hyper-parameters Ridge regression has one hyper-parameter, the λ,that penalizes the $\hat{\beta}_j$ coefficients, i.e., as larger λ is, more costly is to have larger $\hat{\beta}_j$ coefficients. The right value for λ is problem dependent.

Table 8.3 Advantages and disadvantages of ridge regression.

Advantages	Disadvantages
• Strong mathematical foundation • Easily interpretable • Deals better with correlated predictive attributes than ordinary least squares.	• Number of instances must be larger than number of attributes • Sensitive to outliers • Data should be normalized • When relation between predictive and the target attributes is non-linear, uses information poorly

Advantages and disadvantages of ridge regression The advantages and disadvan- tages of ridge regression are shown in Table 8.3.

8.2.3.2 Lasso Regression

The least absolute shrinkage and selection operator (lasso) regression algo- rithm is another penalized regression algorithm, that can deal efficiently with high-dimensional data sets. It performs attribute selection by taking into account not only the predictive performance of the induced model, but also the complexity of the model. The complexity is measured by the number of predictive attributes used by the model. It does this by including in the equation of the multivariate linear regression model an additional weighting term, which depends on the sum of the $\hat{\beta}_j$ weights modules. The weight values define the importance and number of predictive attributes in the induced model.

The lasso algorithm usually produces sparse solutions. Sparse means that a large number of predictive attributes have zero weight, resulting in a regression model that uses a small number of predictive attributes. As

well as attribute selection, the lasso algorithm also performs shrinkage. Mathematically, the lasso algorithm is very well founded.

The lasso formulation is quite similar to the ridge formulation, as presented in Equations (8.14) and (8.13):

$$\underset{\hat{\beta}_0,\ldots,\hat{\beta}_p}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} error(y_i, \hat{y}_i) + \lambda \times \sum_{j=1}^{p} |\hat{\beta}_j| \right\} \tag{8.15}$$

However, they result in substantially different models. This is because the lasso formulation favors the existence of many zero $\hat{\beta}_j$ coefficients. To illustrate why this happens, let us assume that $\hat{\beta}_1 = 0.2$ and $\hat{\beta}_2 = 0.3$. The ridge approach will have $\sum_{j=1}^{p} \hat{\beta}_j^2 = 0.2^2 + 0.3^2 = 0.04 + 0.09 = 0.13$ while the lasso approach will have $\sum_{j=1}^{p} |\hat{\beta}_j| = |0.2| + |0.3| = 0.5$. But if, instead, $\hat{\beta}_1 = 0.5$ and $\hat{\beta}_2 = 0$, ridge regression will have $\sum_{j=1}^{p} \hat{\beta}_j^2 = 0.5^2 + 0^2 = 0.25 + 0 = 0.25$, a value larger than

Table 8.4 Advantages and disadvantages of the lasso.

Advantages	Disadvantages
• Strong mathematical foundation • Easier interpretation than ordinary least squares or ridge regression because it produces simpler models (with fewer predictive attributes) • Deals better with correlated predictive attributes than ridge regression or ordinary least squares; • Automatically discounts irrelevant attributes	• Number of instances must be larger than number of attributes • Sensitive to outliers • Data should be normalized • When the relationship between predictive and target attributes is non-linear, uses information poorly

before, but the lasso will have 0.5 + 0 = 0.5, the same value as before. This example shows that ridge regression promotes shrinkage while lasso promotes attribute selection by setting some of the $\hat{\beta}_j$ weights to 0 but also shrinking some other coefficients.

Assessing and evaluating results As with MLR and ridge regression, the main result of lasso regression are the estimates of the $\hat{\beta}_j$ coefficients. Lasso regres- sion is also a multivariate linear model, but uses a different method to learn the $\hat{\beta}_j$ coefficients.

Setting the hyper-parameters Like ridge regression, lasso regression has one hyper-parameter, λ, which penalizes the $\hat{\beta}_j$ coefficients; that is, the larger λ is, more costly is to have larger $\hat{\beta}_j$ coefficients. The correct value for λ is problem dependent.