Chapter 6: Linear Model Selection and Regularization

Jeremy Selva

Introduction

This chapter introduces alternative fitting procedures can yield better

- Prediction Accuracy
- Model Interpretability

These methods are

- Subset Selection
- Shrinkage
- Dimension Reduction

6.1.1 Best Subset Selection

Fit a separate least squares regression for all possible combination of p predictors and find the best one.

1. Let M_0 be the sample mean. Denote as the *null model* with no predictors.
2. For $k = 1, 2, \ldots p$:
   a) Fit $\binom{p}{k}$ models that contain exactly k predictors.
   b) $M_k =$ the best among these $\binom{p}{k}$ models via largest R^2
3. Choose the optimal model from $M_0, \ldots, M_p$ using a different model selection criteria from R^2.
   - Cross-validation, C_p, AIC, BIC, or Adjusted R^2.

In step 3, we can't use R^2 because it increases monotonically as the number of predictors included in the models increases.

Unfortunately, this method is computationally infeasible for large values of p.

6.1.2 Stepwise Selection

Forward stepwise selection helps to reduce the computational burden by adding predictors one at a time. In each step, we add the predictor that gives the **greatest** improvement to the model.

1. Let M_0 be the sample mean. Denote as the *null model* with no predictors.
2. For $k = 0, 1, \ldots p-1$:
   a) Fit $(p-k)$ models that contain predictors in M_k and one additional predictors not in M_k.
   b) $M_{k+1} =$ the best among these $(p-k)$ models via largest R^2.
3. Choose the optimal model from $M_0, \ldots, M_p$ using a different model selection criteria from R^2.
   - Cross-validation, C_p, AIC, BIC, or Adjusted R^2.

6.1.2 Stepwise Selection

Backward stepwise selection begins with a model containing all predictors and we eliminate the predictors one at a time. In each step, we eliminate the predictor that gives the **least** error to the model when removed.

1. Let M_p be the *full model* with p predictors.
2. For $k = p, p-1, \ldots 1$:

   a) Fit k models that contain all but one predictors in M_k, for a total of $k-1$ predictors.

   b) $M_{k-1} =$ the best among these $(k-1)$ models via largest R^2.
3. Choose the optimal model from $M_0, \ldots, M_p$ using a different model selection criteria from R^2.
   - Cross-validation, C_p, AIC, BIC, or Adjusted R^2.

6.1.3 Choosing the Optimal Model

Recall that in step 3 of all our approaches, we can't use R^2 because it increases monotonically as the number of predictors included in the models increases. The alternative model selection criteria must **penalise** models that uses more predictors.

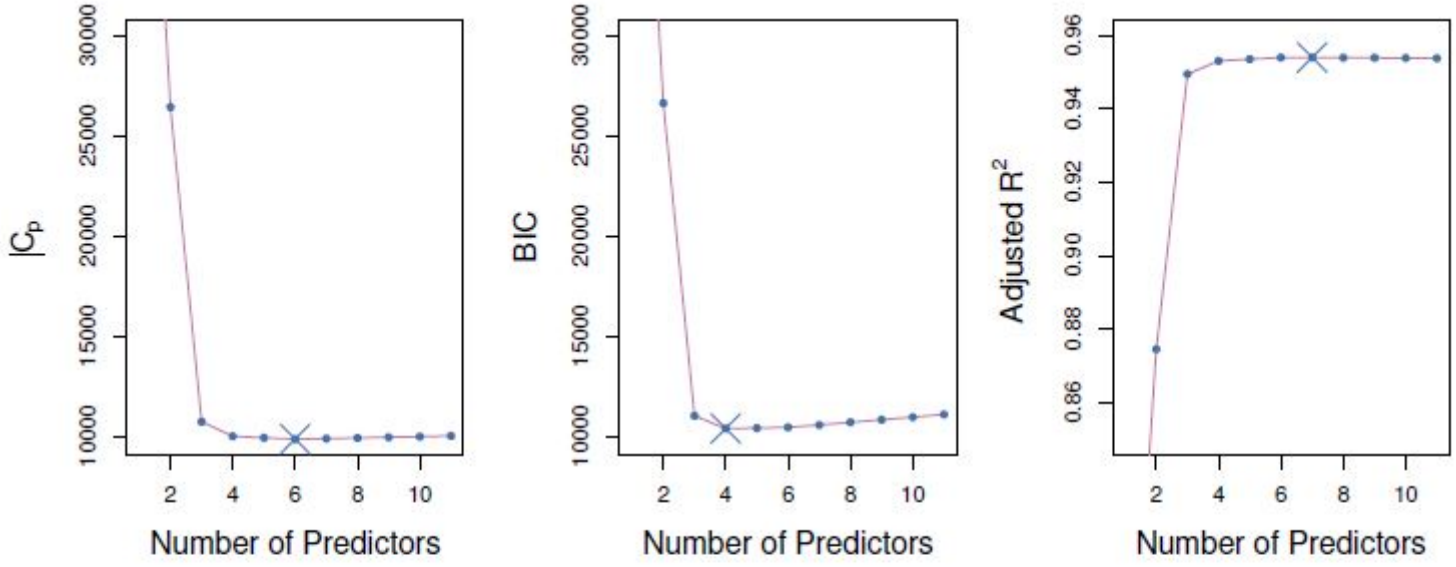


FIGURE 6.2. *C_p, BIC, and adjusted R^2 are shown for the best models of each size for the* `Credit` *data set (the lower frontier in Figure 6.1). C_p and BIC are estimates of test MSE. In the middle plot we see that the BIC estimate of test error shows an increase after four variables are selected. The other two plots are rather flat after four variables are included.*

6.1.3 Choosing the Optimal Model

For a fitted least squares model containing d predictors, where $0 \leq d \leq p$,

$$C_p = \frac{1}{n}\left(\text{RSS} + 2d\hat{\sigma}^2\right)$$

where $\hat{\sigma}^2$ is an estimate of the variance of the error ϵ in model (6.1)

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

In the case of a linear model, $\hat{\sigma}^2 = \text{RSE}^2$ from equation (3.25) in page 80 or

$$\hat{\sigma}^2 = \frac{\text{RSS}}{n - p - 1}$$

The C_p statistic essentially adds **a penalty** of $2d\hat{\sigma}^2$ to the training RSS in order to adjust for the fact that the training error usually underestimate the test error. Hence, we choose the model with the lowest C_p value

6.1.3 Choosing the Optimal Model

For a fitted least squares model containing d predictors, where $0 \leq d \leq p$, given model (6.1)

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

with Gaussian errors like $\epsilon \sim N(0, \sigma^2)$.

Removing irrelevant constants, we have

$$\text{AIC} = C_p = \frac{1}{n}\left(\text{RSS} + 2d\hat{\sigma}^2\right)$$

$$\text{BIC} = \frac{1}{n}\left(\text{RSS} + log(n)d\hat{\sigma}^2\right)$$

Like C_p, we choose the model with the lowest BIC value. However, since $log(n) > 2$ for $n < 7$, the BIC places **a heavier penalty** than C_p and AIC on models with many variables.

6.1.3 Choosing the Optimal Model

For a fitted least squares model containing d predictors, where $0 \leq d \leq p$,

$$Adjusted R^2 = 1 - \frac{\text{RSS}/(n-d-1)}{\text{TSS}/(n-1)}$$

The idea is that **increasing** d, will increase the numerator $\text{RSS}/(n-d-1)$, leading to a **decrease** in the adjusted R^2 value.

Besides using C_p, AIC, BIC, or Adjusted R^2, we can also use the validation set and cross-validation methods discussed in Chapter 5

6.2 Shrinkage

An alternative way to improve the linear model is to fit the model containing all p predictors

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

but **shrinks** the regression coefficients $\beta_0, \beta_1, \ldots, \beta_p$ **towards zero**.

The two best-known techniques are ***ridge regression*** and the ***lasso***.

6.2.1 Ridge Regression

In the least squares fitting procedure, given n samples and p predictors, the regression coefficients $\beta_0, \beta_1, \ldots, \beta_p$ was estimated to minimise the RSS denoted by

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

Ridge regression is similar to least squares, except we are minimizing the regression coefficients by

$$\mathrm{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2$$

where $\lambda \geq 0$ is a ***tuning parameter***.

The second term $\lambda \sum_{j=1}^{p} \beta_j^2$ is called the ***shrinkage penalty***. It can only be small when the regression coefficients $\beta_1, \ldots, \beta_p$ are small. Thus, forcing them to be close to 0. The shrinkage is not applied to the intercept β_0.

6.2.1 Ridge Regression

$$\text{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2$$

λ controls the relative impact of these two terms (at $\lambda = 0$ it is the original least squares estimate). As λ **increase**, the ridge coefficient estimates **shrink towards zero**.

It is best to apply ridge regression after ***standardizing the predictors***, using the formula for each dataset x_{ij}

$$\overline{x_{ij}} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x_j})^2}}$$

This is to ensure each predictor has equal scale/variance/noise and **can be penalised equally**.

A predictor X_k with a variance/noise much larger than others tends to have its corresponding coefficient estimate β_k to be much larger than the others. As the ridge regression penalty involves summing up all the coefficient estimate $\lambda(\sum_{j=1}^{p} \beta_j^2)$, we have a bias as β_k will shrink close to zero at a slower pace as λ increases and the predictor may be mistakenly identified as being useful.

6.2.2 The Lasso

The issue with ***ridge regression*** is its need to include all p predictors. Its ***shrinkage penalty*** $\lambda \sum_{j=1}^{p} \beta_j^2$ can only shrink its regression coefficients $\beta_1, \ldots, \beta_p$ close to 0 but **not exactly** 0 as λ gets large.

This means that it is **hard** to interpret if the most important predictors in the ridge regression are really useful. We still need to **verify** this by creating a new model with just the important predictors. This is be challenging when the number of predictors p gets large.

The ***lasso*** on the other hand, tries to overcome this disadvantage by having its regression coefficients that minisise this quantity

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

where $\lambda \geq 0$ is a ***tuning parameter***.

Lasso's ***shrinkage penalty*** $\lambda \sum_{j=1}^{p} |\beta_j|$ is able to force its regression coefficients to be **exactly** 0 as λ gets large. Thus, able to create models which only contains the useful predictors.

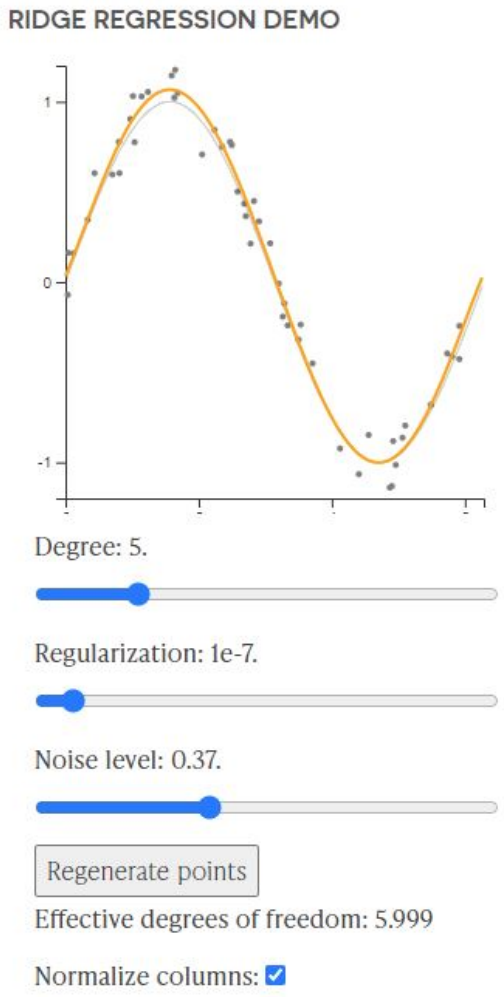
6.2.3 Selecting the Tuning Parameter

Implementing ***ridge regression*** and the ***lasso*** requires a method for selecting a suitable value for the tuning parameter λ.

This is achieved by using cross validation for each value of λ and take the value that gives the least cross-validation error.

The model is then re-fitted using all of the available observations and the selected value of the ***tuning parameter***.

Blog post examples



6.3 Dimension Reduction

The next method involves **transforming** the predictors $X_1, \ldots, X_p$ to $Z_1, \ldots, Z_M$ where $M < p$ and

$$Z_M = \sum_{j=1}^{p} \phi_{jm} X_j$$

for some constants $\phi_{1m}, \phi_{2m}, \ldots, \phi_{pm}, m = 1, \ldots, M$.

The **transformed** variables $Z_1, \ldots, Z_M$ are then used to build a regression model giving these regression coefficients $\theta_0, \ldots, \theta_M$.

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \epsilon_i$$

for $i = 1, . . , n$.

6.3 Dimension Reduction

The condition on the transformation is that it must be possible to express each regression coefficients $\beta_1, \ldots, \beta_p$ of the original predictors $X_1, \ldots, X_p$ in terms of $\theta_0, \ldots, \theta_M$ and in this form

$$\beta_j = \sum_{m=1}^{M} \theta_m \phi_{jm}$$

We will consider two approaches for this task: ***principal components*** and ***partial least squares***

6.3.1 Principal Components Regression

Principal components are projected directions/vector of the data that best represent the data's **variation** or information as close as possible to the data.

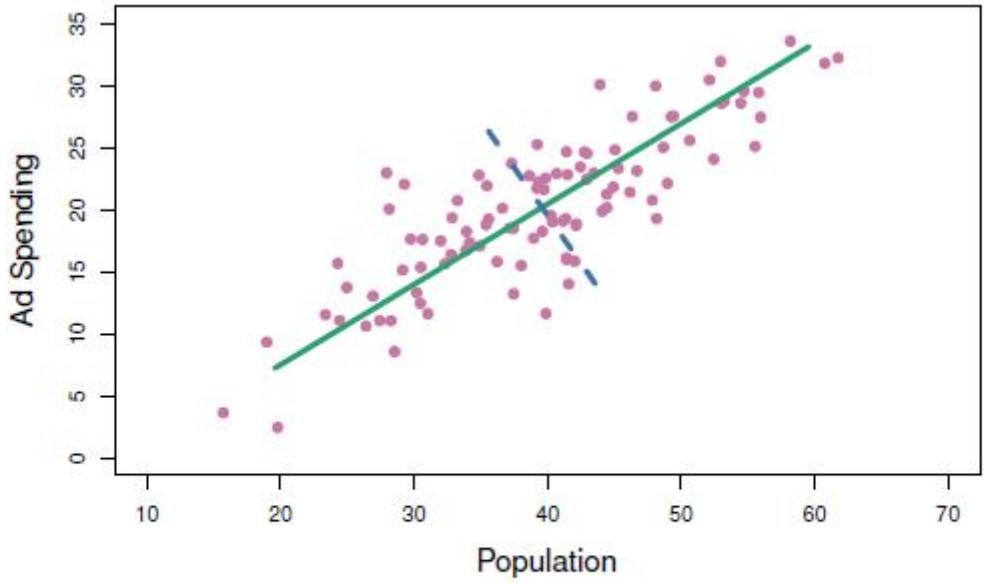


FIGURE 6.14. *The population size (pop) and ad spending (ad) for 100 different cities are shown as purple circles. The green solid line indicates the first principal component, and the blue dashed line indicates the second principal component.*

6.3.1 Principal Components Regression

The first principal component line **minimizes** the sum of the squared perpendicular distances between each point and the green line. These distances are plotted as dashed line segments in the left-hand panel of Figure 6.15

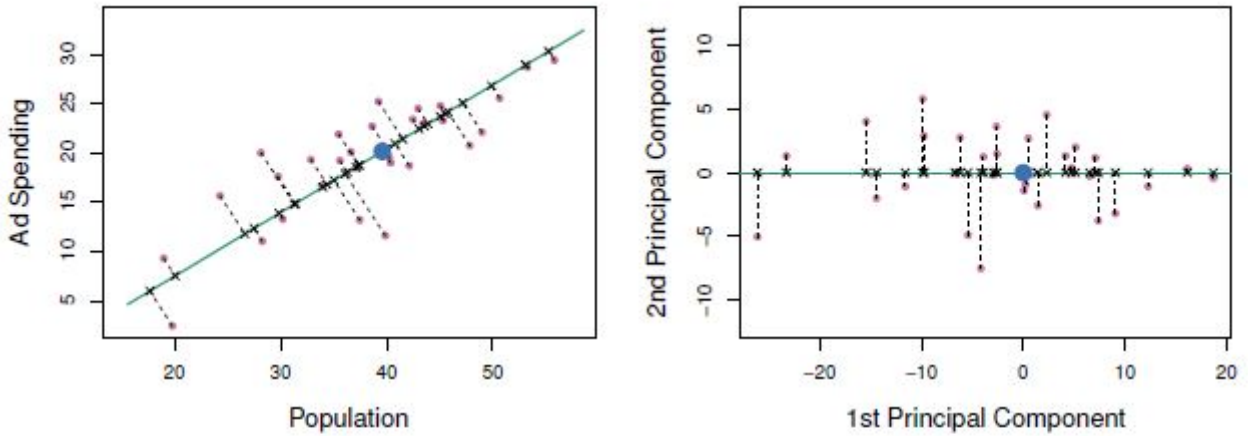


FIGURE 6.15. *A subset of the advertising data. The mean* pop *and* ad *budgets are indicated with a blue circle.* Left: *The first principal component direction is shown in green. It is the dimension along which the data vary the most, and it also defines the line that is closest to all n of the observations. The distances from each observation to the principal component are represented using the black dashed line segments. The blue dot represents* $(\overline{\text{pop}}, \overline{\text{ad}})$. Right: *The left-hand panel has been rotated so that the first principal component direction coincides with the x-axis.*

6.3.1 Principal Components Regression

In terms of formula, the first principal component Z_1 is

$$Z_1 = 0.839 \times (pop - \overline{pop}) + 0.544 \times (ad - \overline{ad})$$

where $\phi_{11} = 0.839, \phi_{21} = 0.544, \overline{pop}$ is the mean of all pop values and $\overline{ad}$ is the mean of all ad values in this data set.

ϕ_{11} and ϕ_{21} were estimated to ensure $Var(\phi_{11} \times (pop - \overline{pop}) + \phi_{21} \times (ad - \overline{ad}))$ is maximised under the condition that $\phi_{11}^2 + \phi_{21}^2 = 1$.

From Z_1, the principal component scores $z_{11}, \ldots, z_{n1}$ can be computed

6.3.1 Principal Components Regression

The values of $z_{11}, \ldots, z_{n1}$ can be seen in the right-hand panel of Figure 6.15.

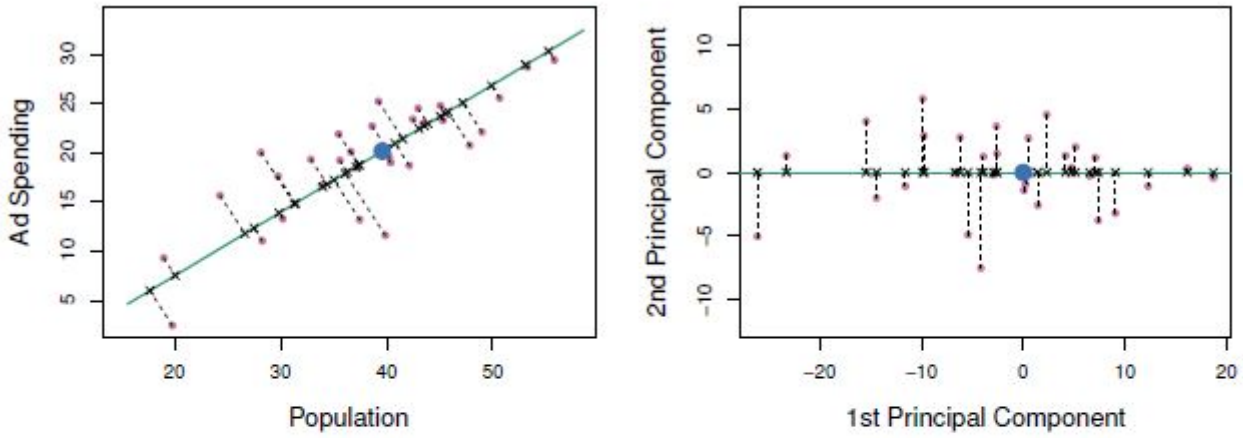


FIGURE 6.15. *A subset of the advertising data. The mean* pop *and* ad *budgets are indicated with a blue circle.* Left: *The first principal component direction is shown in green. It is the dimension along which the data vary the most, and it also defines the line that is closest to all n of the observations. The distances from each observation to the principal component are represented using the black dashed line segments. The blue dot represents* $(\overline{\text{pop}}, \overline{\text{ad}})$. Right: *The left-hand panel has been rotated so that the first principal component direction coincides with the x-axis.*

6.3.1 Principal Components Regression

Figures 6.16 shows how Z_1, the first principal component has a strong positive correlation with pop and ad. Hence, it is able to capture most of the information contained the predictor variable pop and ad.

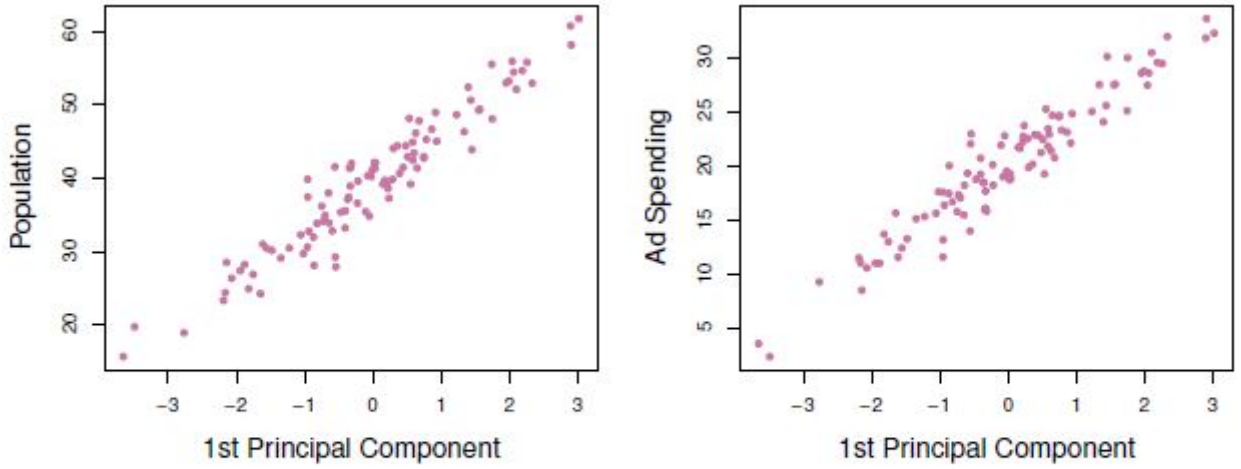


FIGURE 6.16. *Plots of the first principal component scores z_{i1} versus* pop *and* ad. *The relationships are strong.*

6.3.1 Principal Components Regression

The second principal componet Z_2 must obey the following conditions

- It must be expressed in the form $\phi_{12} \times (pop - \overline{pop}) + \phi_{22} \times (ad - \overline{ad})$
- It must be uncorrelated/perpendicular/orthogonal to Z_1 the first principal component.

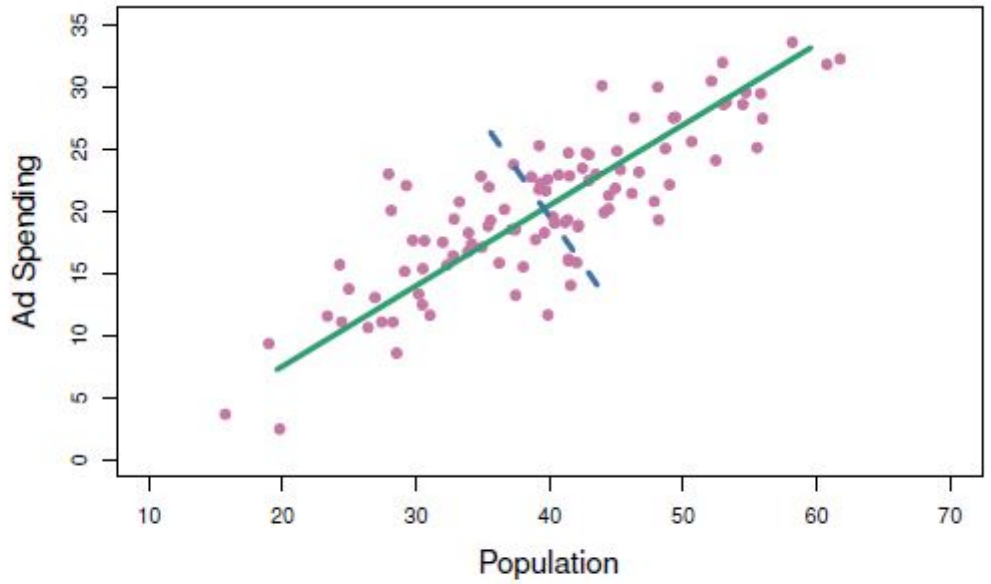


FIGURE 6.14. *The population size (pop) and ad spending (ad) for 100 different cities are shown as purple circles. The green solid line indicates the first principal component, and the blue dashed line indicates the second principal component.*

6.3.1 Principal Components Regression

In terms of formula, the second principal component Z_2 is

$$Z_2 = 0.544 \times (pop - \overline{pop}) - 0.839 \times (ad - \overline{ad})$$

However, this time unlike the first principal component Z_1, its relationship between pop and ad is poor.

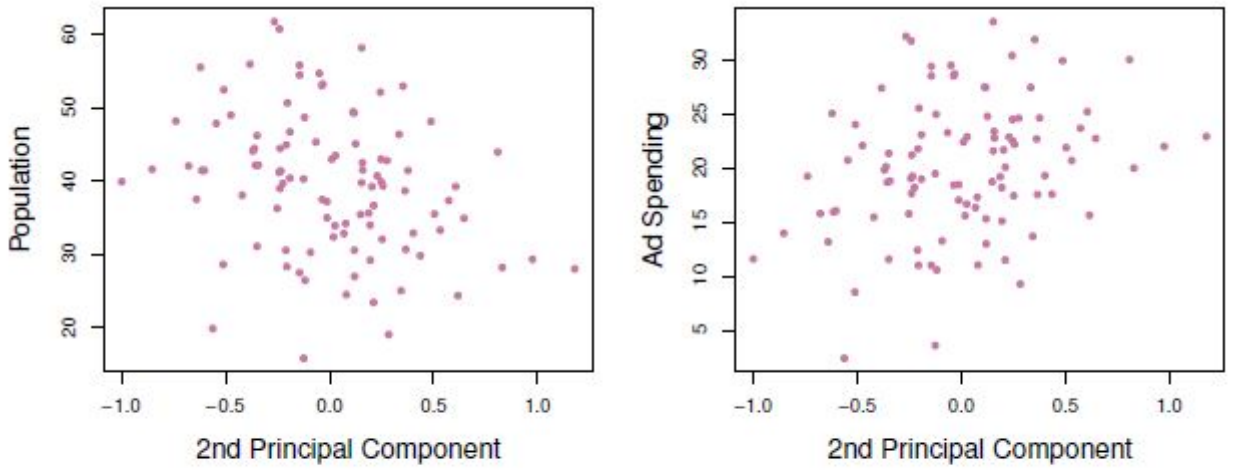


FIGURE 6.17. *Plots of the second principal component scores z_{i2} versus* pop *and* ad. *The relationships are weak.*

6.3.1 Principal Components Regression

Principal components regression **transforms** the predictors $X_1, \ldots, X_p$ into principal components $Z_1, \ldots, Z_M$ where $M < p$ and use these as new predictors to build a regression model giving these regression coefficients $\theta_0, \ldots, \theta_M$.

The assumption made is that

- a small number of principal components is needed to **explain most of the variability** in the data
- principal components has a **good relationship** with the response Y. (which is not always true)

The number of principal components M can be achieved by cross validation starting from 1 to p. The M value that gives the lowest cross validation error will be selected.

Like ridge regression, it is advised to ***standardize the predictors*** before converting them to principal components as high-variance variables will tend to play a larger role in the principal components obtained.

Principal components regression does not work if the principal components does not has a good relationship with the response Y.

6.3.2 Partial Least Squares

The partial Least Squares method overcome this weakness by making use of the **response** Y to **transform** the predictors further.

Like ridge regression, it is advised to ***standardize the predictors*** before converting them to partial least square components.

After standardizing the p predictors, PLS computes the first direction Z_1 as

$$Z_1 = \sum_{j=1}^{p} \phi_{j1} X_j$$

By setting ϕ_{j1} equal to the coefficient B_j from the simple linear regression $Y = intercept + B_j X_j$.

B_j is is proportional to the correlation between Y and X_j. Hence, PLS places the **highest weight** on the variables that are most strongly related to the response.

6.3.2 Partial Least Squares

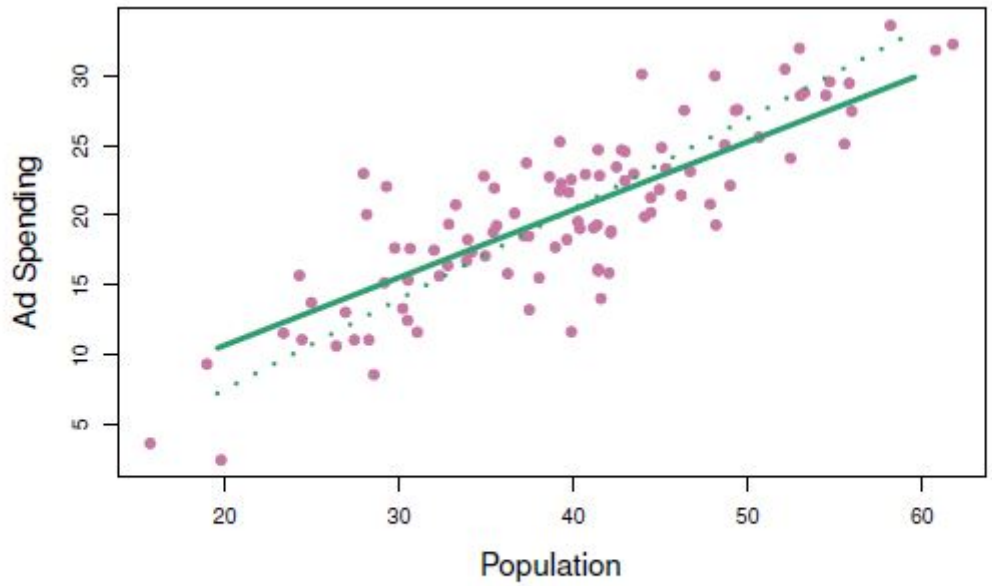


FIGURE 6.21. *For the advertising data, the first PLS direction (solid line) and first PCR direction (dotted line) are shown.*

PLS has chosen a direction that has less change in the ad dimension per unit change in the pop dimension, relative to PCA. This suggests that pop is more highly correlated with the response than is ad.

The PLS direction does not fit the predictors as closely as does PCA, but it does a better job explaining the response.

6.3.2 Partial Least Squares

To get the second PLS direction, each variable X_j is regressed on Z_1 or

$$X_j = intecept + (some\ coefficient)Z_1 + R_j$$

for $j = 1, \ldots, p$, and the residuals R_j are taken. The residuals represent the remaining information that has not been explained by Z_1.

We can then use these residuals to compute Z_2 in the same way that Z_1 was computed on the original data.

$$Z_2 = \sum_{j=1}^{p} \phi_{j2} R_j$$

By setting ϕ_{j2} equal to the coefficient B_j from the simple linear regression $Y = intercept + B_j R_j$.

6.3.2 Partial Least Squares

This iterative approach can be repeated M times to identify multiple PLS components $Z_1, \ldots, Z_M$. Finally, at the end of this procedure, we use least squares to fit a linear model to predict Y using $Z_1, \ldots, Z_M$ in exactly the same fashion as for PCR.

The number of partial least square components M can be achieved by cross validation starting from 1 to p. The M value that gives the lowest cross validation error will be selected.

6.4.1 High-Dimensional Data

Most traditional statistical techniques for regression and classification are intended for low-dimensional setting in which n, the number of observations, is much greater than p, the number of predictors.

However, as technologies improved, it is common to collect an almost unlimited number of feature measurements

We have defined the high-dimensional setting as the case where the number of predictors p **is larger** than the number of observations n.

6.4.2 What Goes Wrong in High Dimensions?

Figure 6.23 further illustrates the risk of carelessly applying least squares when the number of features p is large. Data were simulated with $n = 20$ observations, and regression was performed with between 1 and 20 predictors, each of which was completely unrelated to the response.

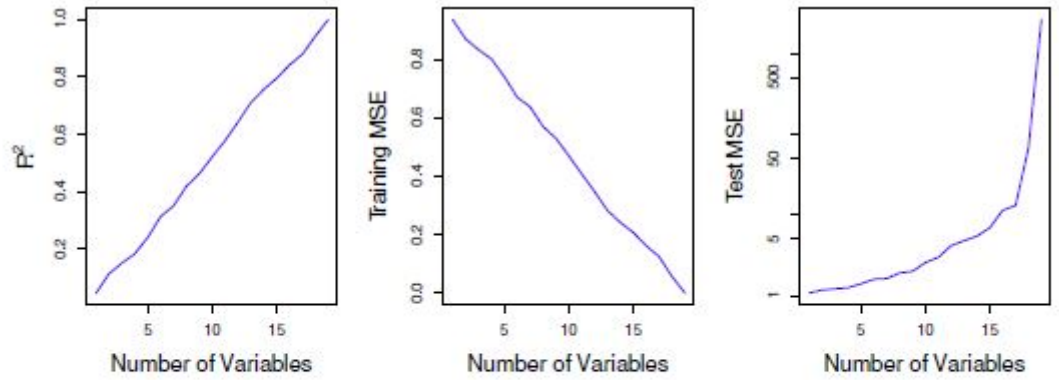


FIGURE 6.23. *On a simulated example with $n = 20$ training observations, features that are completely unrelated to the outcome are added to the model.* Left: *The R^2 increases to 1 as more features are included.* Center: *The training set MSE decreases to 0 as more features are included.* Right: *The test set MSE increases as more features are included.*

6.4.3 Regression in High Dimensions

Figure 6.24 illustrates the performance of the lasso in a simple simulated example of $n = 100$ training observations, and the test mean squared error was evaluated on an independent test set.

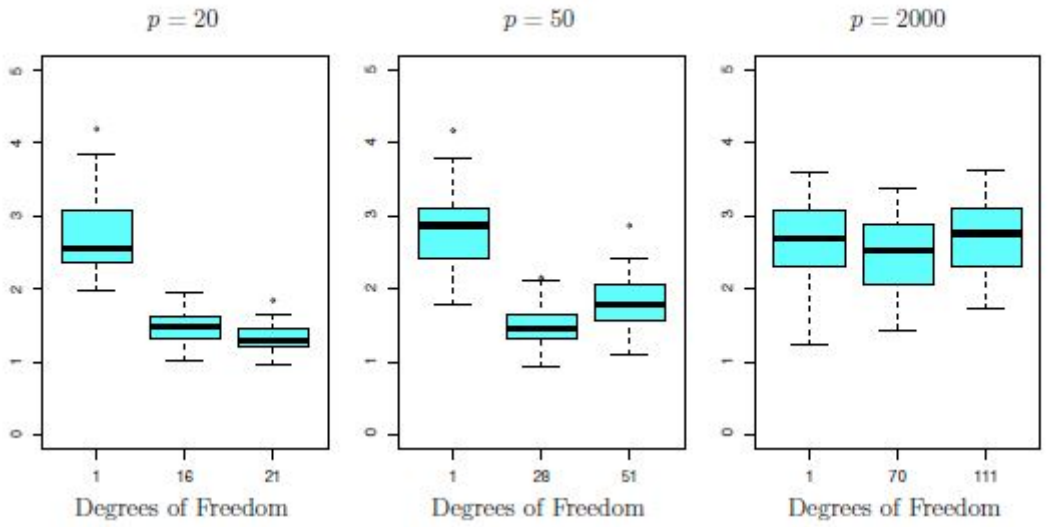


FIGURE 6.24. *The lasso was performed with $n = 100$ observations and three values of p, the number of features. Of the p features, 20 were associated with the response. The boxplots show the test MSEs that result using three different values of the tuning parameter λ in (6.7). For ease of interpretation, rather than reporting λ, the* degrees of freedom *are reported; for the lasso this turns out to be simply the number of estimated non-zero coefficients. When $p = 20$, the lowest test MSE was obtained with the smallest amount of regularization. When $p = 50$, the lowest test MSE was achieved when there is a substantial amount of regularization. When $p = 2{,}000$ the lasso performed poorly regardless of the amount of regularization, due to the fact that only 20 of the 2,000 features truly are associated with the outcome.*

6.4.4 Interpreting Results in High Dimensions

In the high-dimensional setting, it is easy to encounter a predictor that is highly correlated with another predictor. This makes it harder to identify which predictor is meaningful or important.

It is advised to report results on an independent test set, or cross-validation errors.

Models must be further validated on multiple independent data sets.

- For example, suppose a forward stepwise selection results in 17 predictors to predict blood pressure on one independent dataset. It is possible that these 17 predictors may not be selected by the same forward stepwise selection process on another independent dataset.

Models that are effective in prediction, may make use of variables that not be useful in the field of study.