- Instead of least squares fit, other fitting approach can be used to improve prediction accuracy and interpretability.

Predic^n accuracy POV:

- If true rel^n bet^n X and Y is linear, least square estimate will have a low bias.

- If $n >> p$, least squares will produce low variance.
  But if n is not so large, variance will increase $\Rightarrow$ overfitting.

- If $n < p$, variance $= \infty$ $\Rightarrow$ incorrect predic^n.

Interpretability POV:

- In a multiple regression model, some variables are not associated with response var. and hence can be excluded to reduce complexity of model.
- To do this, coef. of corresponding predictors must be set to zero, but least squares approach doesn't predict zero values.

Why consider alternatives to least squares?

- *Prediction Accuracy:* especially when $p > n$, to control the variance.
- *Model Interpretability:* By removing irrelevant features — that is, by setting the corresponding coefficient estimates to zero — we can obtain a model that is more easily interpreted.
- Approaches for automatically performing *feature selection* will be presented.

- Various methods can be used instead of least squares.

- *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.
- *Shrinkage.* This approach involves fitting a model involving all p predictors. However, the estimated coefficients are shrunken towards zero relative to the least squares estimates. This shrinkage (also known as *regularization*) has the effect of reducing variance. Depending on what type of shrinkage is performed, some of the coefficients may be estimated to be exactly zero. Hence, shrinkage methods can also perform variable selection.
- *Dimension Reduction.* This approach involves *projecting* the p predictors into an M-dimensional subspace, where $M < p$. This is achieved by computing M different *linear combinations*, or *projections*, of the variables. Then these M projections are used as predictors to fit a linear regression model by least squares.

① Subset Selection:
② Shrinkage (Regularization)
③ Dimension Reduction

✡ Subset Selection :

Algorithm 6.1 *Best subset selection*

1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, \ldots p$:
   (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.
   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

$^pC_k = \frac{p!}{k! \times (p-k)!}$

- Backward Selection : when $n > p$
  Forward Selection : when $p > n$

- Model containing all predictors will always have smallest RSS and largest R^2. (As these are related to training error)

- In order to choose the best model, we need to select one with min 'test error'.
  ∴ RSS and R^2 cannot be used to predict the best model.

Solution : We can use methods to adjust the training error to account for bias due to overfitting.

Example- Credit data set

For each possible model containing a subset of the ten predictors in the Credit data set, the RSS and R^2 are displayed. The red frontier tracks the best model for a given number of predictors, according to RSS and R^2. Though the data set contains only ten predictors, the x-axis ranges from 1 to 11, since one of the variables is categorical and takes on three values, leading to the creation of two dummy variables

Extensions to other models

- Although we have presented best subset selection here for least squares regression, the same ideas apply to other types of models, such as logistic regression.
- The *deviance*— negative two times the maximized log-likelihood— plays the role of RSS for a broader class of models.

C_p :

Consider 'd' predictors selected during step selection, then:

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

where, $\hat{\sigma}^2$ = variance of error term ε

- C_p statistic adds a penalty of $2d\hat{\sigma}^2$ to training RSS to adjust training error's underestimation of test error.
- Low test error ⇒ low C_p

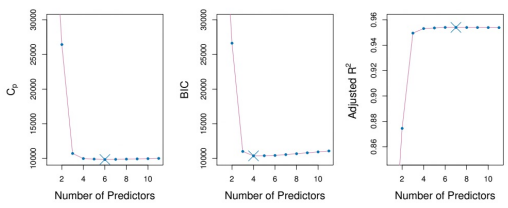


FIGURE 6.2. *C_p, BIC, and adjusted R^2 are shown for the best models of each size for the Credit data set (the lower frontier in Figure 6.1). C_p and BIC are estimates of test MSE. In the middle plot we see that the BIC estimate of test error shows an increase after four variables are selected. The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

AIC (Akaike Information Criterion)

$$AIC = -2\log L + 2d$$

where, L = maximized value of likelyhood function of model.

- In the case of the linear model with Gaussian errors, maximum likelihood and least squares are the same thing, and C_p and AIC are equivalent.

BIC (Binomial Information Criterion)

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

- Like C_p, the BIC will tend to take on a small value for a model with a low test error, and so generally we select the model that has the lowest BIC value.
- Notice that BIC replaces the $2d\hat{\sigma}^2$ used by C_p with a $\log(n)d\hat{\sigma}^2$ term, where n is the number of observations.
- Since log n > 2 for any n > 7, the BIC statistic generally places a heavier penalty on models with many variables, and hence results in the selection of smaller models than C_p.

Adjusted R^2 :

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

Unlike *Cp*, AIC, and BIC, for which a small value indicates a model with a low test error, a large value of adjusted R^2 indicates a model with a small test error. Remember:

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

Maximizing the adjusted R^2 is equivalent to minimizing RSS/(*n*-*d*-1).

While RSS always decreases as the number of variables in the model increases, RSS/(*n*-*d*-1) may increase or decrease, due to the presence of *d* in the denominator.

Unlike the R^2 statistic, the adjusted R^2 statistic pays a price for the inclusion of unnecessary variables in the model.

✡ Validation and Cross-Validation :

- Compute validation error or cross validation error for each model and select best model.
- Better compared to C_p, AIC, BIC as it gives direct test errors.
- Can also be used to choose degrees of freedom (best).

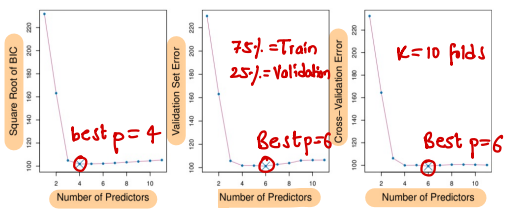


FIGURE 6.3. *For the* Credit *data set, three quantities are displayed for the best model containing d predictors, for d ranging from* 1 *to* 11. *The overall* best *model, based on each of these quantities, is shown as a blue cross.* Left: *Square root of BIC.* Center: *Validation set errors.* Right: *Cross-validation errors.*

✡ Shrinkage Methods :

- We can choose all 'p' predictors and use methods which regularize the coef. estimates.
- Shrinking (regularizing) the estimates, reduces the variance.

○ Shrinking Techniques :

a) **Ridge Regression :**

In Linear Regression, we choose $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2 \dots \hat{\beta}_p$ which minimize

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

Ridge Regression is similar to least squares, except estimating a different quantity.

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

where $\lambda \geq 0$ (Tuning Parameter)

- The tuning parameter serves to control the relative impact of these two terms on the regression coefficient estimates.
- Selecting a good value for λ is critical; cross-validation is used for this.

○ **NOTE :** β_0 is not regularized.

$\lambda \sum_{j=1}^{p} \beta_j^2$ = Shrinkage Penalty

↓

Small when $\beta_1, \beta_2 \dots \beta_p$ is close to zero.

Clearly $\lambda = 0 \Rightarrow$ Ridge coef. estimates = Least Square est.

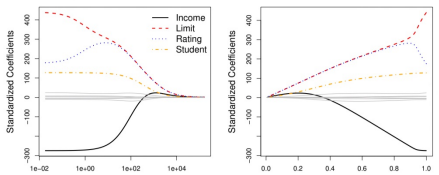


FIGURE 6.4. *The standardized ridge regression coefficients are displayed for the* Credit *data set, as a function of* λ *and* $\|\hat{\beta}_\lambda^R\|_2/\|\hat{\beta}\|_2$.

$\|\beta\|_2 = l_2$ norm of vector

$\|\beta_2\| = \sqrt{\sum_{j=1}^{p} \beta_j^2}$

(It measures dist. of β from 0)

- $\lambda = 0 \Rightarrow$ Least Square Coef.
- As $\lambda \uparrow$, estimates shrink to zero.
- If λ is extremely large, estimates $\to 0$
  - NULL MODEL (No predictors, $p = 0$)

- As $\lambda \uparrow$, l_2 norm of $\hat{\beta}_\lambda^R \downarrow$
  $\Rightarrow$ as $\lambda \uparrow$, $\|\hat{\beta}_\lambda^R\|_2 / \|\hat{\beta}\|$ will $\downarrow$

$\|\hat{\beta}_\lambda^R\|_2 / \|\hat{\beta}\| \in [1, 0]$

- The standard least squares coefficient estimates are *scale equivariant*: multiplying X_j by a constant c simply leads to a scaling of the least squares coefficient estimates by a factor of $1/c$.
- In other words, regardless of how the jth predictor is scaled, $X_j\hat{\beta}_j$ will remain the same.

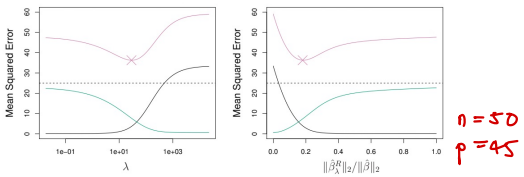
- In contrast, the ridge regression coefficient estimates can change *substantially* when multiplying a given predictor by a constant, due to the sum of squared coefficients term in the penalty part of the ridge regression objective function.

- Therefore, it is best to apply ridge regression after *standardizing the predictors*, using the formula

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}}$$

- Practical note: try raw, standardized, and normalized data.

- Denominator = Std. deviation of "standardized" predictor.
- As predictors are standardized, std. error = 1



n = 50
p = 45

FIGURE 6.5. *Squared bias (black), variance (green), and test mean squared error (purple) for the ridge regression predictions on a simulated data set, as a function of λ and $\|\hat{\beta}_\lambda^R\|_2/\|\hat{\beta}\|_2$. The horizontal dashed lines indicate the minimum possible MSE. The purple crosses indicate the ridge regression models for which the MSE is smallest.*

- Ridge regression does have one obvious disadvantage: unlike subset selection, which will generally select models that involve just a subset of the variables, ridge regression will include all *p* predictors in the final model

+ The penalty $\lambda \sum_{j=1}^{p} \beta_j^2$ will shrink all coefficients to zero (unless $\lambda = \infty$)
+ This will not hamper accuracy but interpretability as p becomes large.

✡ Lasso Regression :

- Lasso coef. $\hat{\beta}_\lambda^L$, minimize the quantity

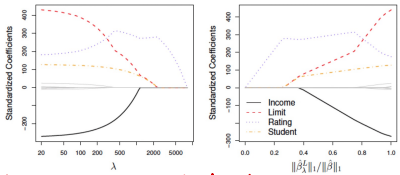
$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| = RSS + \lambda \sum_{j=1}^{p} |\beta_j|$$

- Lasso Penalty = $|\beta_j|$
- Lasso uses L_1 penalty
- L_1 norm of coefficient vector β is given by:

$$\|\beta\|_1 = \sum |\beta_j|$$

- NOTE: Unlike Ridge reg., Lasso reg. can have some coef. estimates exactly equal to zero when λ is large.

∴ Lass is more interpretable than Ridge.



Right: Initially, only 'rating' is part of the model. As $\lambda \uparrow$ 'limit' appears, then student and finally 'rating'.

But in case of Ridge reg. all predictors are considered irrespective of the value of λ.

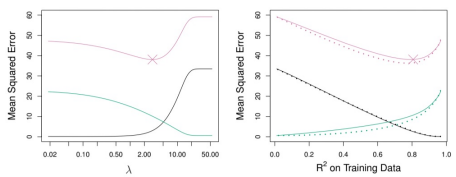


FIGURE 6.8. Left: *Plots of squared bias (black), variance (green), and test MSE (purple) for the lasso on a simulated data set.* Right: *Comparison of squared bias, variance, and test MSE between lasso (solid) and ridge (dotted). Both are plotted against their R^2 on the training data, as a common form of indexing. The crosses in both plots indicate the lasso model for which the MSE is smallest.*

+ It can be clearly seen from the right plot that Lasso gives less MSE than Ridge Regression

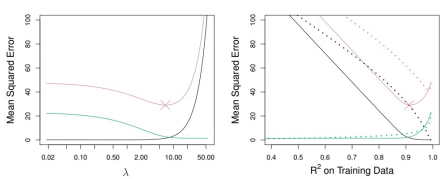


FIGURE 6.9. Left: *Plots of squared bias (black), variance (green), and test MSE (purple) for the lasso. The simulated data is similar to that in Figure 6.8, except that now only two predictors are related to the response.* Right: *Comparison of squared bias, variance, and test MSE between lasso (solid) and ridge (dotted). Both are plotted against their R^2 on the training data, as a common form of indexing. The crosses in both plots indicate the lasso model for which the MSE is smallest.*

+ From the right plot, MSE (min) for Ridge reg. is slightly smaller than Lasso.

+ Ridge reg. performs better than Lasso in terms of prediction error (accuracy).

✱ Selecting the tuning parameter (λ) :

- Cross validation is used for selecting best λ.

Algo.
1) Grid of λ values is chosen.
2) Cross validation is performed for each λ.
3) Select the λ with minimum cross-validation error.
4) Model is refit using all observations and selected (best) λ.

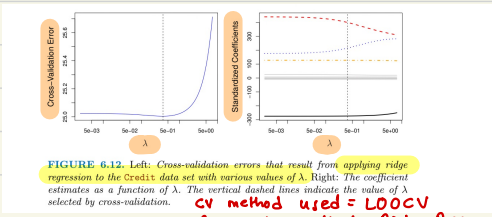


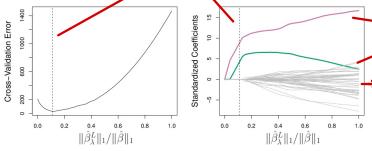
FIGURE 6.12. Left: *Cross-validation errors that result from applying ridge regression to the* Credit *data set with various values of* λ. Right: *The coefficient estimates as a function of* λ. *The vertical dashed lines indicate the value of* λ *selected by cross-validation.*

CV method used = LOOCV
Regression method = Ridge Reg.

- **Inference :**

① Many λs can give same error.
② In this case, we can use least squares solution.

Min. CV error



Related variables ($p = 2$) aka signal var
Grey : unrelated predictors aka noise var.

FIGURE 6.13. Left: *Ten-fold cross-validation MSE for the lasso, applied to the sparse simulated data set from Figure 6.9.* Right: *The corresponding lasso coefficient estimates are displayed. The two signal variables are shown in color, and the noise variables are in gray. The vertical dashed lines indicate the lasso fit for which the cross-validation error is smallest.*

CV method = 10 fold CV
Regression method = Lasso

-

✡ Elastic Net :

Elastic Net

- The elastic net is a regularization method that combines L1 and L2 regularizations, and enforces the coefficients of correlated variables to vary together as λ changes. The Elastic Net penalty is:

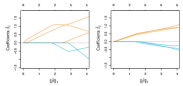
$$\lambda\left[\frac{1}{2}(1-\alpha)\|\beta\|_2^2+\alpha\|\beta\|_1\right]$$

where $\alpha \in [0,1]$ is a parameter that can be varied. When $\alpha = 1$, it reduces to the L1-norm or lasso penalty, and with $\alpha = 0$, it reduces to the squared L2-norm, corresponding to the ridge penalty.

89

Elastic Net

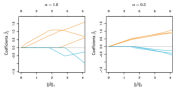
- Example: Six variables, highly correlated in groups of three. The lasso estimates ($\alpha = 1$), as shown in the left panel, exhibit somewhat erratic behavior as the regularization parameter λ is varied. In the right panel, the elastic net with ($\alpha = 0.3$) includes all the variables, and the correlated groups are pulled together.



90

Elastic Net

- Of course, this example is idealized, and in practice the group structure will not be so cleanly evident. But by adding some component of the ridge penalty to the L1-penalty, the elastic net automatically controls for strong within-group correlations.



91

Elastic Net vs. Lasso Constraints

- Figure 4.2 compares the constraint region for the elastic net (left image) to that of the lasso (right image) when there are three variables. We see that the elastic-net ball shares attributes of the L2 ball and the L1 ball; the sharp corners and edges encourage selection, and the curved contours encourage sharing of coefficients.

92

✡ Dimension Reduction Methods :

Instead of choosing a subset of variables (predictors) or shrinking the coef., we can transform the predictors and then fit a model on these transformed variables.

↳ This is k/a Dimensionality Reduction.

Dimension Reduction Methods: details

- Let $Z_1, Z_2, \ldots, Z_M$ represent $M < p$ linear combinations of our original p predictors. That is,

$$\left\{ Z_m = \sum_{j=1}^{p} \phi_{mj} X_j \quad (1) \right\}$$

for some constants $\phi_{m1}, \phi_{m2}, \ldots, \phi_{mp}$.

- We can then fit the linear regression model,

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \epsilon_i, \quad i = 1, \ldots, n, \quad (2)$$

using ordinary least squares.

Regression coef. = $\theta_0, \theta_1, \theta_2 \ldots \theta_M$

Dim. Red^n : Reducing $(p+1)$ coef. to $(M+1)$ coef. where $M < P$.

$$z_m = \sum_{j=1}^{p} \phi_{jm} x_j$$

$$\sum_{m=1}^{M} \theta_m z_{im} = \sum_{m=1}^{M} \theta_m \sum_{j=1}^{p} \phi_{jm} x_{ij}$$

$$= \sum_{j=1}^{p} \sum_{m=1}^{M} \theta_m \phi_{jm} x_{ij}$$

$$= \sum_{j=1}^{p} \beta_j x_{ij}$$

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

- Dim. Red^n methods work in 2 steps:

  Step 1 : Transform p predictors into M transformed predictors. $Z_1, Z_2, Z_3 \ldots Z_m$.

  Step 2 : Fit the model using these transformed predictors.

  How we perform Step 1 is important.

- Transformation techniques:
  a) Principle Component
  b) Partial Least Squares

a) **Principle Component Regression:** (Principle Component Analysis - PCA)

- We have a data matrix $X = (n \times p)$ → predictors (columns); ↓ datapoints (rows)
- First Principle Component direction:
  Direc^n along which observations vary the most. (Maximum Variance)

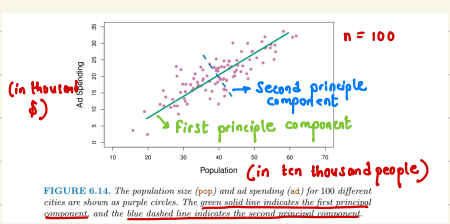


FIGURE 6.14. *The population size (pop) and ad spending (ad) for 100 different cities are shown as purple circles. The green solid line indicates the first principal component, and the blue dashed line indicates the second principal component.*

Component Loadings. (ϕ_{11} and ϕ_{21})

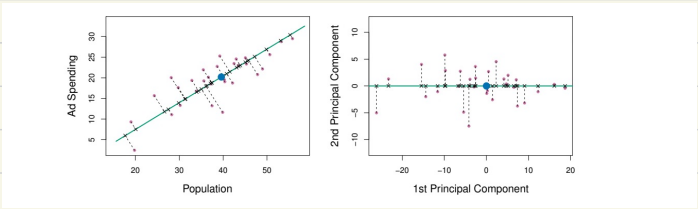
$Z_1 = 0.839\,(pop - \overline{pop}) + 0.554\,(ad - \overline{ad})$

$\text{slope} = \frac{0.554}{0.839}$ → means

Out of every possible combination, such that $\phi_{11}^2 + \phi_{21}^2 = 1$, the above combina^n gives max. variance.

- First Principle Component direction (line):
  If we project all the observations along this line, we will get maximum variance.

NOTE - Projecting a point on a line means finding a point on the line such that line passing thru these 2 points is ⊥ to first principle component line.



- Principle Component Score :

$$z_{i1} = 0.839\,(pop_i - \overline{pop}) + 0.544\,(ad_i - \overline{ad})$$

$z_{11}, z_{21}, z_{31} \ldots z_{n1}$ ($n = 100$ in this ex.) are k/a principle component scores.

- First Principle Component Line is closest to data.

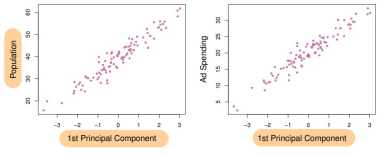


FIGURE 6.16. *Plots of the first principal component scores z_{i1} versus* pop *and* ad. *The relationships are strong.*

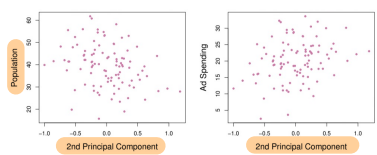


FIGURE 6.17. *Plots of the second principal component scores z_{i2} versus* pop *and* ad. *The relationships are weak.*

- PCA involves choosing first M principle components $Z_1, Z_2 \ldots Z_M$.
- Fit a linear regression model with least squares.

As more princ. comp. are added to the model, bias ↓ and variance ↑.
This results in U-shape of MSE in the curve.

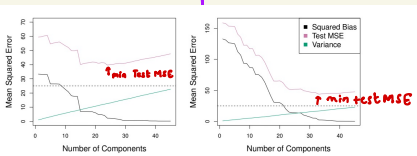


FIGURE 6.18. *PCR was applied to two simulated data sets. In each panel, the horizontal dashed line represents the irreducible error.* Left: *Simulated data from Figure 6.8.* Right: *Simulated data from Figure 6.9.*

PCR (PCA) will do well when first few principle comp. will capture most of the variation in the data.

+ PCR and Lasso, are very closely related.

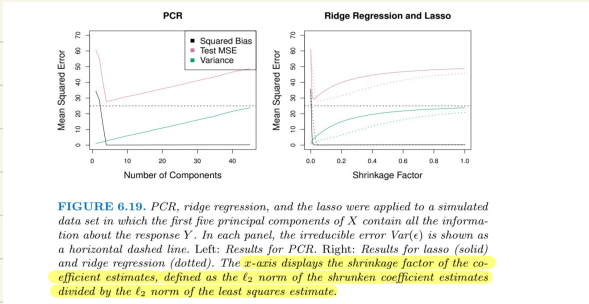


FIGURE 6.19. *PCR, ridge regression, and the lasso were applied to a simulated data set in which the first five principal components of X contain all the information about the response Y. In each panel, the irreducible error Var(ε) is shown as a horizontal dashed line. Left: Results for PCR. Right: Results for lasso (solid) and ridge regression (dotted). The x-axis displays the shrinkage factor of the coefficient estimates, defined as the ℓ_2 norm of the shrunken coefficient estimates divided by the ℓ_2 norm of the least squares estimate.*

- It is recommended to 'standardize' each predictor before applying PCR.
- Standardization ensures all variables are on same scale.
- If standardization is not performed:
  Variables (predictors) with large variance will tend to have larger impact on PCR.
- It units are same, standardization can be avoided.

- PCR identifies linear combinations, or *directions*, that best represent the predictors $X_1, \ldots, X_p$.
- These directions are identified in an *unsupervised* way, since the response Y is not used to help determine the principal component directions.
- Consequently, PCR suffers from a potentially serious drawback: there is no guarantee that the directions that best explain the predictors will also be the best directions to use for predicting the response. Imp.

PLS (Partial Least Squares)

Data where $p >> n$ = High Dimensional Data

Partial Least Squares: (PLS)

- Similar to PCR in:
  1. Selecting set of features $Z_1, Z_2, \ldots, Z_M$ that are linear combination of original features.
  2. Fit a linear model on these new features using least squares.

- PLS identifies these features using SUPERVISED approach.
- It uses response variable Y to find new features which approximate old features and also related to response.

- PLS helps in finding directions that help explain both response and predictors

- After standardizing the p predictors, PLS computes the first direction Z_1 by setting each φ_{1j} equal to the coefficient from the simple linear regression of Y onto X_j.
- One can show that this coefficient is proportional to the correlation between Y and X_j. Hence, in computing $Z_1 = \sum_{j=1}^{p} \phi_{1j} X_j$, PLS places the highest weight on the variables that are most strongly related to the response.
- Subsequent directions are found by taking residuals and then repeating the above prescription.

+ Predictors and response is standardized before PLS.
+ PLS can reduce bias but increases variance, hence useful.

+ If $p >> n$, model will overfit.
+ $R^2 \longrightarrow 1$, Train MSE will decrease.
+ Test MSE will increase
+ ∴ Few features should be selected which approximate the response variable.

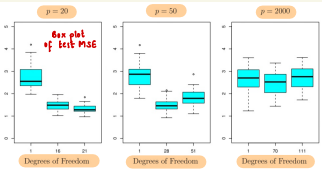


FIGURE 6.24. *The lasso was performed with $n = 100$ observations and three values of p, the number of features. Of the p features, 20 were associated with the response. The boxplots show the test MSEs that result using three different values of the tuning parameter λ in (6.7). For ease of interpretation, 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.*

lasso fit. Figure 6.24 highlights three important points: (1) regularization or shrinkage plays a key role in high-dimensional problems, (2) appropriate tuning parameter selection is crucial for good predictive performance, and (3) the test error tends to increase as the dimensionality of the problem (i.e. the number of features or predictors) increases, unless the additional features are truly associated with the response.