Chapter 4

Linear Methods for Regression

In these notes we introduce a couple of linear methods similar to regression but that are designed to improve prediction not for interpreting parameters. We will introduce the singular value decomposition and principal component analysis. Both these concept will be useful throughout the class.

4.1 Linear Predictors

Before computers became fast, linear regression was almost the only way of attacking certain prediction problems. To see why, consider a model such as this

$$Y = \beta_0 + \beta_1 e^{\beta_2 X} + \epsilon, \tag{4.1}$$

finding the βs that minimize, for example, least squares is not straight forward. A grid search would require many computations because we are minimizing over a 3-dimensional space.

Technical note: For minimizing least squares in (4.1) the Newton-Raphson algorithm would work rather well. But we still don't get an answer in closed form.

As mentioned, the least square solution to the linear regression model:

$$Y = \beta_0 + \sum_{j=1}^{p} X_j \beta_j + \epsilon$$

has a closed form *linear solution*. In Linear Algebra notation we write: $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, with $\boldsymbol{\beta} \equiv (\beta_0, \beta_1, \beta_2)'$. The important point here is that for any set of predictors x the prediction can be written as a linear combination of the observed data $\hat{Y} = \sum_{i=1}^{n} w_i(x) y_i$. The $w_i(x)$ are determined by the X_js and do not depend of $\mathbf{y}$.

What is the prediction $\hat{Y}$ for x ?

When we say linear regression we do not necessarily mean that we model the Y as an actual line. All we mean is that the expected value of Y is a linear combination of predictors. For example, if λ is a fixed quantity (i.e. we are not estimating it) then this is a linear model:

$$Y = \beta_0 + \beta_1 \sin(\lambda X) + \beta_2 \cos(\lambda X) + \epsilon.$$

To see this simply define $X_1 = \sin(\lambda X)$ and $X_2 = \cos(\lambda X)$.

For the model (4.1) defined above, we can not do the same because $X_1 = e_2^{\beta} X$ contains a parameter.

If the linear regression model holds then the least squares solution has various nice properties. For example, if the ϵs are normally distributed then $\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimate and is normally distributed as well. Estimating the variance components is simple: $(\mathbf{X}'\mathbf{X})^{-1}\sigma^2$ with σ^2 the error variance, var(ϵ). σ^2 is usually well estimated using the residual sum of squares.

If the ϵs are independent and identically distributed (IID) then $\hat{\boldsymbol{\beta}}$ is the linear unbiased estimate with the smallest variance. This is called the Gauss-Markov theorem.

Technical note: Linear regression also has a nice geometrical interpretation. The prediction is the orthogonal projection of the vector defined by the data to the *hyper-plane* defined by the regression model. We also see that the least squares estimates can be obtained by using the Graham Schmidt algorithm which orthogonalized the covariates and then uses simple projections. This algorithm also helps us understand the QR decomposition. For more see the book for more details. More latter.

4.1.1 Testing Hypotheses

Review the t-test, i.e. estimate divided by standard estimate idea... and what happens when we have to estimate the standard error.

The fact that we can get variance estimates from regression, permits us to test for simple hypotheses. For example

$$\hat{\beta}_j / \mathbf{se}(\hat{\beta}_j) \sim t_{N-p-1}$$

under the assumptions of normality for ϵ. When ϵ is not normal but IID, then the above is asymptotically normal.

If we want to test significance of various coefficients. In this case we can generalize to the F-test.

$$\frac{(RSS_0 - RSS_1)/(p_1 - p_0)}{RSS_1/(N - p_1 - 1)}$$

Under normality assumptions this statistic (the F-statistic) follows a $F_{p_1-p_0}, N - p_0 - p_1$ distribution.

Similarly we can form confidence intervals (or balls). For the case of multiple coefficients we can use the fact that

$$\frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \mathbf{X}' \mathbf{X} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\hat{\sigma}^2}$$

Follow a χ^2_{p+1} distribution

4.2 Graham Schmidt

One can show that the regression coefficient for the jth predictor is the simple regression coefficient of y on this predictor adjusted for all others (obtained using Graham-Schmidt).

For the simple regression problem (with not intercept)

$$Y = X\beta + \epsilon$$

the least square estimate is

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

Can you see for the constant model?

Mathematicians write the above solution as

$$\hat{\beta} = \frac{< \mathbf{x}, \mathbf{y} >}{< \mathbf{x}, \mathbf{x} >}$$

We will call this operation regressing $\mathbf{y}$ on $\mathbf{x}$ (its the projection of $\mathbf{y}$ onto the space spanned by $\mathbf{x}$)

The residual can then be written as

$$\mathbf{r} = \mathbf{y} - \beta\mathbf{x}$$

What was the solution for β_1 to $Y = \beta_0 + \beta_1 X$?

We can rewrite the result as

$$\hat{\beta}_1 = \frac{< \mathbf{x} - \bar{x}\mathbf{1}, \mathbf{y} >}{< \mathbf{x} - \bar{x}\mathbf{1}, \mathbf{x} - \bar{x}\mathbf{1} >}$$

Notice that we the Graham-Schmidt permits us to estimate the β_j in a multivariate regression problem by first regressing $\mathbf{x}_j$ on $\mathbf{x}_1$, then the residuals from that on bx_2, up to $\mathbf{x}_p$. Then regressing $\mathbf{y}$ on the final residuals.

Notice that if the $\mathbf{x}$s are correlated then each predictor affects the coefficient of the others.

The interpretation is that the coefficient of a predictor $\hat{\beta}_j$ is the regression of $\mathbf{y}$ on $\mathbf{x}_j$ after $\mathbf{x}_j$ has be *adjusted* for all other predictors.

4.3 Subset Selection and Coefficient Shrinkage

When we have many predictors and the linear model seems to fit the data well, there are still reasons why we are not satisfied: Prediction accuracy and interpretation.

Two ways that we can achieve this is by

1. Using a subset of the predictors (notice more predictors always means less bias) and

2. Shrinking some of the coefficients toward zero

Ridge regression permits us to do 2.

4.3.1 Subset Selection

Although the least squares estimate is the linear unbiased estimate with minimum variance, it is possible that a biased estimate will give us a better mean squared error.

Consider a case where the true model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

and that X_1 and X_2 are almost perfectly correlated (statisticians say X_1 and X_2 are co-linear). What happens if we leave X_2 out?

Then the model is very well approximated by

$$Y = \beta_0 + (\beta_1 + \beta_2) X_1 + \epsilon$$

and we may get a good estimate of Y estimating 2 parameters instead of 3. Our estimate will be a bit biased but we may lower our variance considerably creating an estimate with smaller EPE than the least squares estimate.

We wont be able to interpret the estimated parameter, but out prediction may be good.

In Figure 4.3.1 we demonstrate results from fitting various subset models to a simulated example where the true model is a linear model with 10 predictors and we observe 15 outcomes.

In subset selection regression we select a number of covariates to include in the model. Then we look at all possible combinations of covariates and pick the

one with the smallest RSS. In Figure 4.3.1 we show for each value of covariates included a point representing the RSS for a particular model. We do this for a training and test set.

As expected, for the training set the RSS becomes smaller as we include more predictors. However, for the training data using a biased model produces better results.

To see this in real data let us consider the prostate cancer data set presented in the HTF book. (The data set is available from the lasso2 R package)

These data come from a study that examined the correlation between the level of prostate specific antigen and a number of clinical measures in men who were about to receive a radical prostatectomy. It is data frame with 97 rows and 9 columns.

The predictors available in the data set are: lcavol, log(cancer volume), lweight, log(prostate weight)age, agelbph, log(benign prostatic hyperplasia amount) svi, seminal vesicle invasionlcp, log(capsular penetration)gleason, Gleason scorepgg45, and percentage Gleason scores 4 or 5lpsa.

The data is described in more detail here:

Stamey, T.A., Kabalin, J.N., McNeal, J.E., Johnstone, I.M., Freiha, F., Redwine, E.A. and Yang, N. (1989) Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate: II. radical prostatectomy treated patients, *Journal of Urology* **141** (5), 1076–1083.

Notice that smaller models tend to do better.

Figure 4.1: Prediction error (RSS) for all possible subset models for training and test sets. The solid lines denote the minimums.

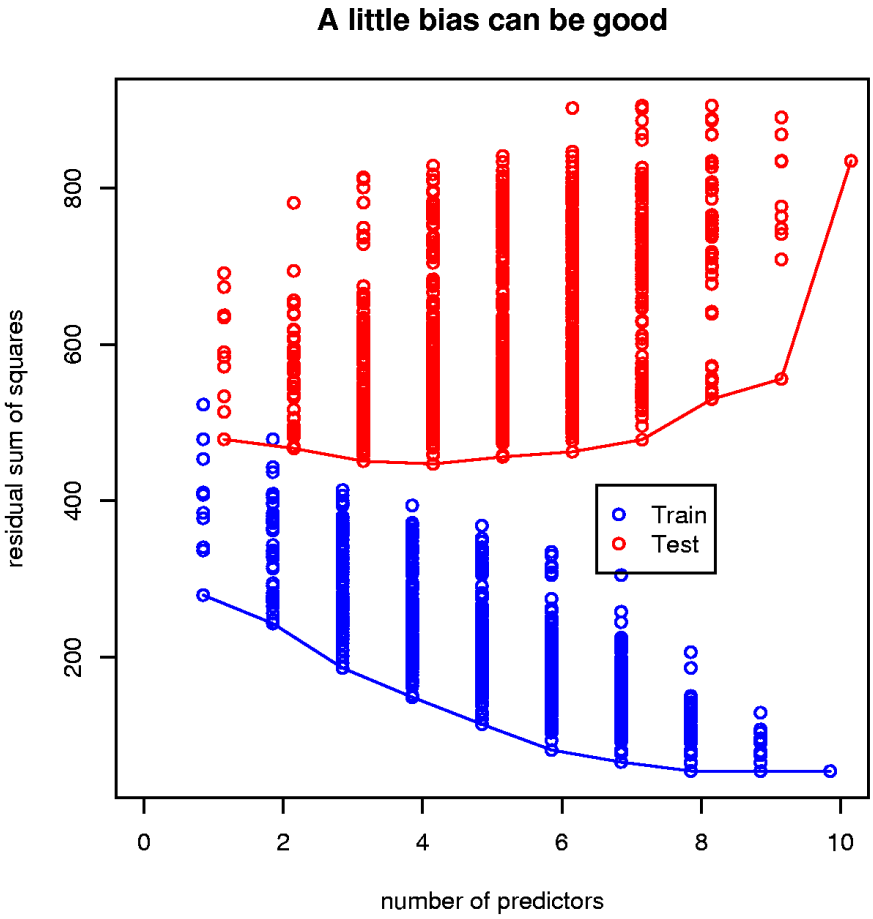
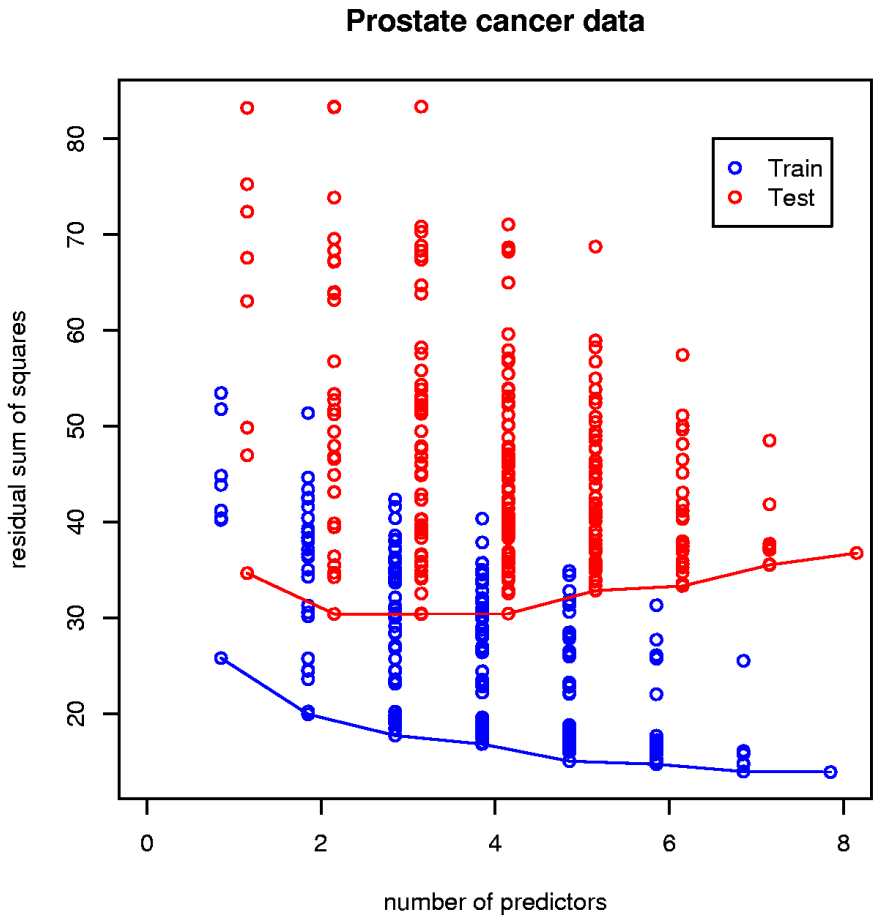


Figure 4.2: Prediction error (RSS) for all possible subset models for training and test sets for prostate cancer data. The solid lines denote the minimums.



For a given number of predictors, how do we find the model that gives the smallest RSS? There are algorithms that do this, but you do not really want to use this. Better things are about to be described.

How do we choose the number of covariates to include? Thats a bit harder.

4.3.2 Shrinkage methods

By only considering some of the covariates we were able to improve our prediction error. However, the choice of one covariate over an another can sometimes be a very arbitrary decision as including either works well but both together do no work as well (this happens often with correlated predictors).

We can think of the subset selection procedure as one that *shrinks* some of the coefficient to 0. But what if we do this in a smoother way? Instead of either keeping it (multiply by 1) or not (multiply by 0), let's permit smoother shrinkage (multiply by a number between 0 and 1).

4.3.3 Ridge Regression

For ridge regression instead of minimizing least squares we *penalize* for having to many β that are big by considering the following minimization criteria:

$$\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2.$$

We will denote the parameter vector that minimizes this $\hat{\boldsymbol{\beta}}^{\text{ridge}}$. Here λ is a penalty sometimes called the *complexity parameter*.

One can demonstrate mathematically that minimizing the above expression is equivalent to minimizing the regular RSS

$$\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \text{ subject to } \sum_{j=1}^{p} \beta_j^2 \leq s$$

where s is inversely proportional to lambda.

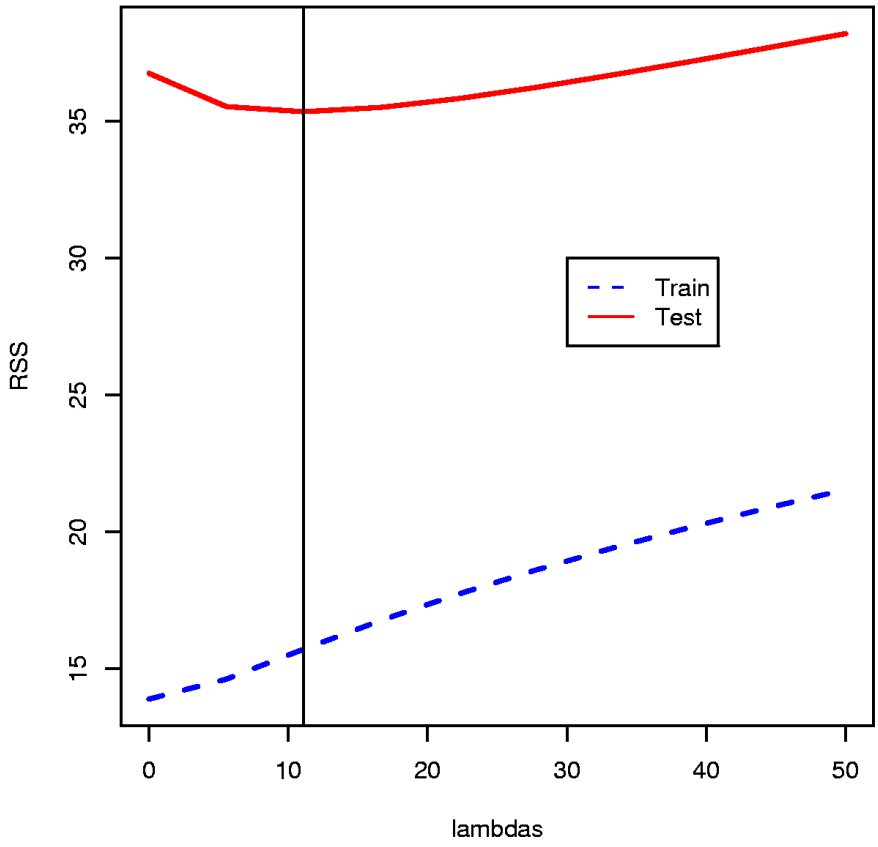
Notice that when λ is 0, we get the least squares estimate. However, as λ gets bigger, over fitting gets more expensive as larger values of β penalize the criterion more. The smallest penalty occurs when all the βs are 0. This gives us an estimate with small variance but likely large bias.

Although this problems looks complicated it turns out the resulting predictor is a linear estimate!

One can show that the solution is (in linear algebra notation)

$$\hat{\boldsymbol{\beta}}^{\text{ridge}} = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}$$

Figure 4.3: Prediction error (RSS) for ridge regression with varying penalty parameters



In Figure 4.3.3 we see the RSS in a test and training set for the prostate cancer data for various values of λ.

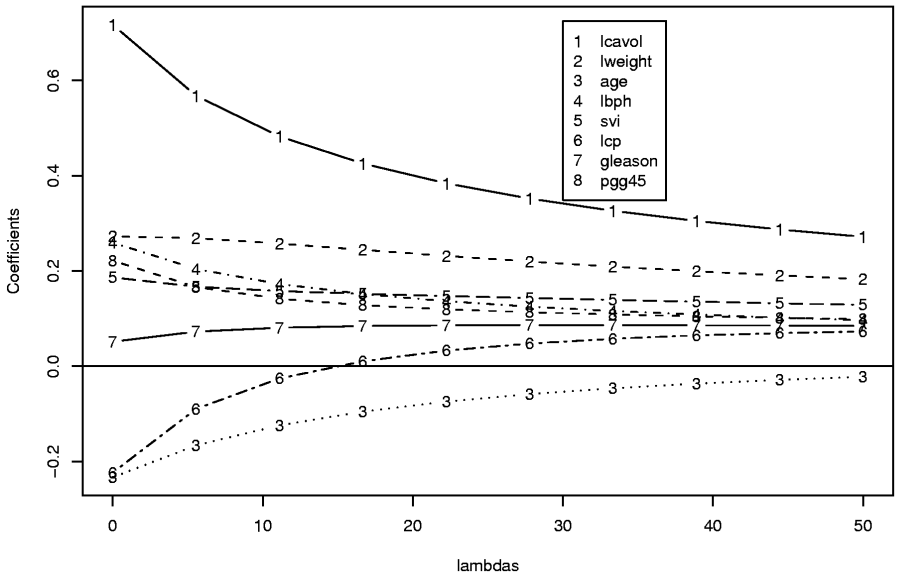
As expected the RSS in the training set is best when $\lambda = 0$ (no shrinkage, nothing stopping us from over-fitting). However, for the training set the smallest RSS occurs for $\lambda \approx 10$

The least squares estimates are given below. Notice age has a significant protective effect. This is at odds with out intuition.

	Est	SEt	Pr(>|t|)	
(Intercept)	-0.10	1.42	0.9434	
lcavol	0.59	0.10	9.58e-07	***
lweight	0.73	0.28	0.0160	*
age	-0.03	0.01	0.0257	*
lbph	0.18	0.07	0.0244	*
svi	0.50	0.32	0.1329	
lcp	-0.16	0.10	0.1299	
gleason	0.07	0.17	0.6983	
pgg45	0.01	0.004	0.1199	

Ridge regression shrinks the regression coefficients toward 0. Notice what happens to these coefficients as λ grows. Notice in particular what happens to age.

Figure 4.4: Estimated coefficients using ridge regression with various penalty parameters.



4.3.4 SVD and PCA

The singular value decomposition (SVD) of the centered input matrix $\mathbf{X}$ gives us insight into the nature of ridge regression.

This decomposition is extremely useful in many statistical analysis methods. We will see it again later.

The SVD of an $N \times p$ matrix $\mathbf{X}$ is

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}'$$

with $\mathbf{U}$ and $\mathbf{V}$ $N \times p$ and $p \times p$ orthogonal matrices and $\mathbf{D}$ a $p \times p$ diagonal matrix with entries $d_1 \geq d_2 \geq \ldots d_p \geq 0$ called the singular values of $\mathbf{X}$.

Technical Note: $\mathbf{U}$ is an orthogonal basis for the space defined by the columns of $\mathbf{X}$ and $\mathbf{V}$ is an orthogonal basis for the space defined by the rows of $\mathbf{X}$.

We can show that the least squares predictor for linear regression is

$$\begin{aligned} \hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}^{ls} &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \mathbf{U}\mathbf{U}'\mathbf{y} \end{aligned}$$

Technical Note: $\mathbf{U}'\mathbf{y}$ are the coordinates of $\mathbf{y}$ with respect to the orthogonal basis $\mathbf{U}$

The ridge solution can be expressed as

$$\mathbf{X}\hat{\boldsymbol{\beta}}^{\text{ridge}} = \mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'y$$

$$\begin{aligned} &= \mathbf{U}\mathbf{D}(\mathbf{D}^2 + \lambda I)^{-1}\mathbf{D}\mathbf{U}'y \\ &= \sum_{j=1}^{p} \mathbf{u}_j \frac{d_j^2}{d_j^2 + \lambda} \mathbf{u}_j' \mathbf{y} \end{aligned}$$

Notice that because $\lambda > 0$, $\frac{d_j^2}{d_j^2+\lambda} \leq 1$. Like linear regression, ridge regression computes the coordinates of $\mathbf{y}$ with respect to the orthogonal basis $\mathbf{U}$. It then shrinks these coordinates by the factors $\frac{d_j^2}{d_j^2+\lambda}$. This means that a greater amount of shrinkage occurs when λ is big and for smaller d_js.

What does having a small d_j represent? A way to understand this is by looking at the *principal components* of the variables in $\mathbf{X}$.

If the $\mathbf{X}$ are centered, the sample covariance matrix is given by $\mathbf{X}'\mathbf{X}/N$ and $\mathbf{X}'\mathbf{X}/$ can be written as

$$\mathbf{X}'\mathbf{X} = \mathbf{V}\mathbf{D}^2\mathbf{V}.$$

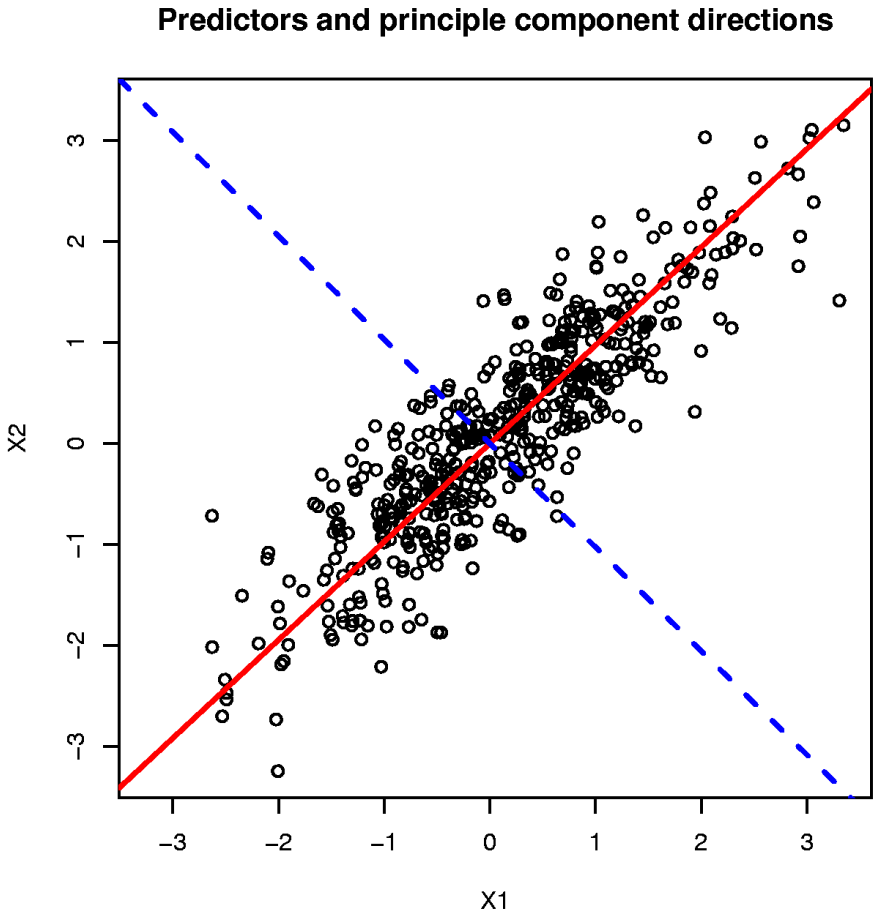
Technical note: this is the eigen decomposition of $\mathbf{X}'\mathbf{X}$.

The v_js are called the eigen values and also the principal components directions of $\mathbf{X}$. Figure 4.3.4 shows a scatterplot of $\mathbf{X}$ and the directions as red (solid) and blue (dashed) lines.

The first principal component $\mathbf{z}_1 = \mathbf{X}v_1$ has the property that it has the largest sample covariance among all normalized (coefficients squared add up to 1) linear combinations of $\mathbf{X}$. The sample variance is d_1^2/N.

The derived variable $\mathbf{z}_1 = \mathbf{X}v_1 = \mathbf{u}_1 d_1$ is called the first principal component.

Figure 4.5: Plot of two predictors, X_2 versus X_1, and the principal component directions



Similarly $\mathbf{z}_j = \mathbf{X}v_1$ is called the jth principal component. $\mathbf{XV} = \mathbf{UD}$ is a matrix with principal components in the columns. Figure 4.3.4 shows these.

We now see that ridge regression shrinks coefficients related to principal components with small variance. This makes sense because we have less information about his.

In the case of Figure 4.3.4, we can think of it as weight and height, we are saying predict with the sum and ignore the difference. In this case, the sum give much more info than the difference.

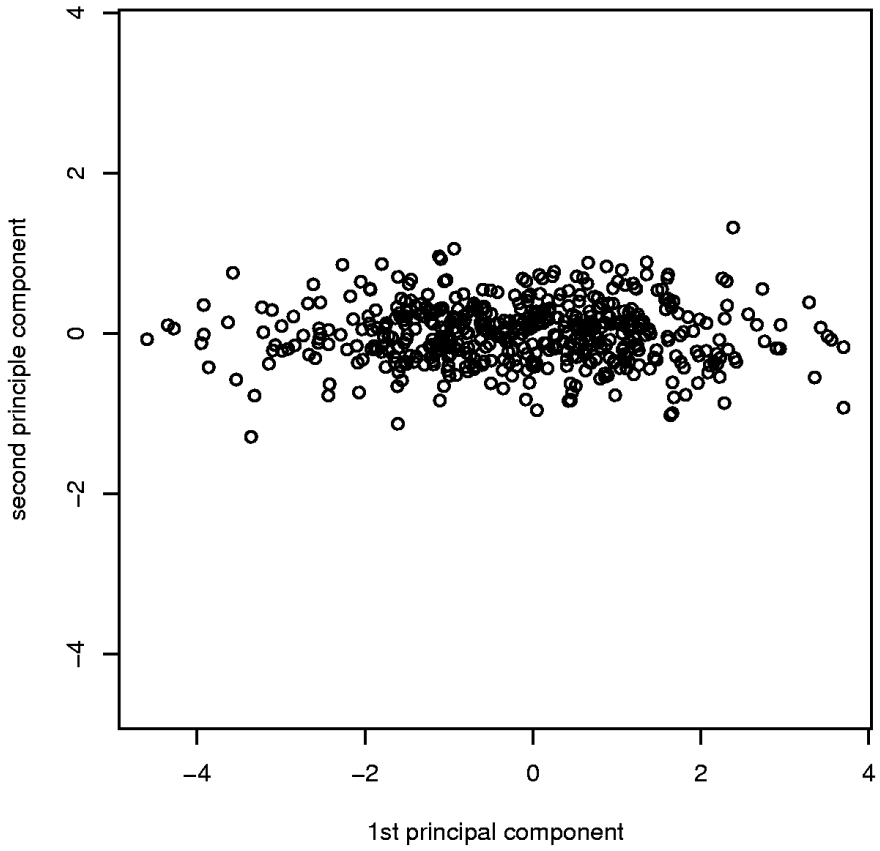
Principal Component Regression

Principal component regression disregard the need to interpret coefficients as effects sizes of the covariates. Instead we include the linear combinations of covariates that contain the most information.

We regress the outcome Y on a few principal components $\mathbf{z}_1, \ldots, \mathbf{z}_M$. Notice that if $M = p$ we are back to simply doing regression on all the covariates. However, for principal components with very small eigenvalues d_j it makes little sense to include them in the regression because it means that the subjects differ very little in z_j thus we have little information to learn about the effect of this component.

Because the principal components are orthogonal, the prediction is simply

Figure 4.6: Plot of two principal components of $\mathbf{X}$.



$$\mathbf{y}^{PCR} = \sum_{m=1}^{M} \hat{\theta}_m \mathbf{z}_m$$

with

$$\hat{\theta}_m = \sum_{i=1}^{N} z_{i,m} y_i$$

Partial least squares is a method, similar to principal component regression, that chooses the components to include based on the correlation with Y. This makes the method non-linear and thus computationally less attractive. The fact that in practice the method ends up being almost identical to principal component regression makes it even less attractive.

4.3.5 The Lasso

The lasso's definition is similar to that of ridge regression. However, we obtain very different results.

$$\sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq s$$

The lasso does not provide a linear algorithm

In practice one sees that the lasso makes more coefficients 0. This is sometimes nicer for interpret-ability. See the book and papers on lasso for more information.

Other methods we will not discuss here are principal component regression and partial least squares regression.