Handout 13

Ridge Regression

Consider the car price data which contain information on the cars sold in the U.S. in 1993. The response variable is price and there are 6 other variable - city MPG, hwy MPG, engine size, HP, tank size and weight. The goal is to relate the response variable Y(price) to the other 6 variables. A preliminary investigation via matrix plot (only a few independent variables are taken for better visualization) show that some of the variables need to be transformed. We have transformed price, X_1, X_2 and X_3 by natural logarithm and X_4 by a square root. We will continue to call them variable, Y, X_1, X_2 etc. The correlation matrix shows that not only is Y well related to independent variables, but the independent variables are well correlated among themselves (also called multicollnearity). The method of ridge regression is well suited for such a data set.

The regression model can be expressed in the Gauss-Markov framework $Y = X\beta + \varepsilon$, where X is $n \times p$, $E(\varepsilon) = 0$ and $Cov(\varepsilon) = \sigma^2 I$. The least squares estimate of β is obtained by solving the normal equations $X^T X\beta = X^T Y$ and the solution is $\hat{\beta} = (X'X)^{-1}X'Y$. Now if $R = X^T X$ is singular, that matrix R is not invertible and any computing packages will return an error when trying to carry out a least squares method for such a case. The situation is not all that different if R is nearly singular. Note that R is singular if $p > n$, i.e., there are more beta parameters (to be estimated) than the number of observations. Singularity or near singularity can also occur if $p < n$, but the independent variables are highly correlated among themselves (also called multicollnearity). How do we handle such cases? There are many methods and some of the well known ones are - ridge regression, partial least squares, stepwise regression and principal components regression. We will only discuss here the first two methods. It should be pointed out that these two methods work well if there is a substantial amount of multicollinearity. If the independent variables are nearly uncorrelated among themselves, i.e., when R is nearly a diagonal matrix, these methods may not perform well especially when p/n is not small.

What is the ridge regression method? Let us recall that the least squares estimate is given by $\hat{\beta} = R^{-1}X^T Y$ with $E(\hat{\beta}) = \beta$ and $Cov(\hat{\beta}) = \sigma^2 R^{-1}$. When R is nearly singular, the length $||\hat{\beta}||$ of $\hat{\beta}$ can be very large (to be discussed below) with non-negligible probability and this will happen even if $||\beta||$ is not large. Ridge regression tries to solve this problem by making sure that the estimate of β is not large in length. This is achieved by carrying out a modification of the least squares method, whereby the minimization of $||Y - X\beta||^2$ with respect to β is carried out only among those β's which are not large in length.

Ridge estimate

For this method, one minimizes the modified criterion

$$||Y - X\beta||^2 + k||\beta||^2, k > 0,$$

with respect to β. Note that when $k = 0$, this leads to the ordinary least squares estimate of β. When $k > 0$, the penalty term $k||\beta||^2$ in the above the criterion function encourages looking for solutions for which $||\beta||$ is not too large. As a matter of fact, if k is very large ($k \approx \infty$), it will nearly force all the β parameters to be zero or close to zero. A important point about scaling of variables needs to be made here. If all the independent variables are nearly equally scaled then the penalty term given above is meaningful. However if the scales are quite different, then the penalty term should be $k\beta^T D\beta$, where D is a diagonal matrix with diagonal entries proportional to the variances of the independent variables. In this note we will assume that the **independent variables have been standardized so that they are all equally scaled**.

Before we proceed further, note that the minimization of the modified criterion leads to new normal equations and the solution $\hat{\beta}(k)$ is fairly simple:

$$(X^T X + kI)\beta = X^T Y,\ \hat{\beta}(k) = (X^T X + kI)^{-1} X^T Y.$$

Note that $\hat{\beta}(0) = (X^T X)^{-1} X^T Y$, the least squares estimate and $\hat{\beta}(\infty) = 0$.

Choice of k.

In order to carry out the ridge method successfully, one needs to select the penalty constant k. there are many criteria for selecting k and a few are listed here. Technical details are given in the Appendix. We will begin with generalized cross-validation criterion (GCV). If a criterion is minimized at $k = \hat{k}_{opt}$, then $\hat{k}_{opt}$ is to be used for the ridge regression. Though $\hat{k}_{opt}$ values for different criteria may be different, but they are usually quite close to each other .

GCV criterion: Let $SSE(k) = ||Y - X\hat{\beta}(k)||^2$ be the residual sum of squares using the ridge estimate and let $H(k) = X(X^T X + kI)^{-1} X^T$ be the modified hat-matrix. The GCV criterion is

$$GCV(k) = SSE(k)/[1 - tr(H(k))/n].$$

Akaike-type criterion: This criterion is given by

$$AKA(k) = SSE(k) + 2\hat{\sigma}^2(k) tr(H(k)), \text{ where}$$
$$\hat{\sigma}^2(k) = SSE(k)/tr[(I - H(k))^2]$$

Mallows-type criterion: It is similar to Akaike-type criterion, except that $\hat{\sigma}^2(k)$ is replaced by $\hat{\sigma}^2(k_0)$ where $k_0 > 0$ is small, not too far away from zero.

Cross-validation: Cross-validation criterion (also called leave-one-out cross-validation) method consists in obtaining repeated estimates of β each time deleting one of the n observations. Let $\hat{\beta}^{(-i)}(k)$ be the ridge estimate of β by deleting the i^{th} case with the penalty parameter k, i.e., $\hat{\beta}^{(-i)}(k)$ is obtained by running a ridge regression by deleting i^{th} rows of X and Y. Then the cross-validated estimate of the prediction error is obtained by

$$CV(k) = \sum (Y_i - x_i^T \hat{\beta}^{(-i)}(k))^2,$$

where x_i^T is the i^{th} row of X. It may seem that the CV procedure is computationally expensive since it involves estimating n estimates of β for each k. Fortunately, it turns out that there is no need for such extensive computations. As a matter of fact it is enough to calculate β only once for each k. It turns out that

$$Y_i - x_i^T \hat{\beta}^{(-i)}(k) = (Y_i - x_i^T \hat{\beta}(k))/(1 - h_{ii}),$$

where h_{ii} is the i^{th} diagonal element of $H(k)$. Hence the cross-validataion criterion can be simplified to

$$CV(k) = \sum (Y_i - x_i^T \hat{\beta}(k))^2/(1-h_{ii})^2.$$

If one replaces h_{ii}'s by their average $\bar{h} = n^{-1}\sum h_{ii}$, then one is lead to the GCV criterion since $\bar{h} = n^{-1}\sum h_{ii} = n^{-1}tr(H(k))$.

Estimation of σ^2

In order to estimate we may choose a value of k that is quite a bit smaller that $\hat{k}_{opt}$. Denote $X\beta$ by μ. It can be shown that

$$E(SSE(k)) = \sigma^2 tr[(I - H(k))^2] + ||\mu - \mu(k)||^2 := variance + bias^2,$$

where $\mu(k) = X\beta(k)$, $\beta(k) = E[\hat{\beta}(k)] = (R + kI)^{-1}\mu$. Note that smaller the value of k, smaller is the bias. So for a value of $k > 0$ that is quite a bit smaller than k_{opt}, the bias term becomes negligible and we can get a reasonable estimate of σ^2 as

$$\hat{\sigma}^2 = SSE(k)/tr[(I - H(k))^2].$$

It will be shown later that

$$Cov(\hat{\beta}(k)) = \sigma^2 (R + kI)^{-1} R (R + kI)^{-1}.$$

So an estimate of $Cov(\hat{\beta}(\hat{k}_{opt}))$ is

$$s^2(\hat{\beta}(\hat{k}_{opt})) = \hat{\sigma}^2 (R + \hat{k}_{opt})^{-1} R (R + \hat{k}_{opt})^{-1}.$$

Data: 1993 cars.

We have transformed and standardized the variables, but we still call them Y, X_1, .. Here is the correlation matrix

	Y	X_1	X_2	X_3	X_4	X_5	X_6
Y	1.00						
X_1	-0.765	1					
X_2	-0.677	0.935	1				
X_3	0.733	-0.824	-0.702	1			
X_4	0.842	-0.753	-0.661	0.769	1		
X_5	0.731	-0.858	-0.812	0.796	0.746	1	
X_6	0.771	-0.883	-0.828	0.897	0.780	0.894	1

It is quite clear from the correlation matrix that there is a strong multicollinearity. As a confirmation of that we have obtained the eigenvalues of $X^T X$ and they are

464.87, 37.55, 23.06, 16.18, 6.65, 3.68.

Note that the first eigenvalue alone explains 84% of the total variability (i.e., 464.87/(sum of eigenvalues)=0.84).

The regression $Y = \beta_1 X_1 + \cdots + \beta_6 X_6 + \varepsilon$ is fitted and the coefficients are

Coef	Estimate	Std.Error	t value	p-val
X_1	-0.3648	0.2042	-1.787	0.0775
X_2	0.1495	0.1673	0.894	0.3740
X_3	-0.0745	0.1367	-0.545	0.5872
X_4	0.5745	0.0909	6.322	0.0000
X_5	-0.0040	0.1269	-0.032	0.9249
X_6	0.1948	0.1777	1.096	0.2759

Residual standard error: 0.5095 on 87 degrees of freedom

Multiple R-squared: 0.7545, Adjusted R-squared: 0.7376

F-statistic: 44.57 on 6 and 87 DF, p-value: < 2.2e-16

The stepwise method with AIC criterion chose variables X_1 and X_4 and here are the results

Coefficients:

Coef	Estimate	Std.Error	t value	p-val
X_1	-0.3029	0.08028	-3.772	0.0003
X_4	0.6141	0.08028	7.649	0.0000

Residual standard error: 0.5066 on 90 degrees of freedom

Multiple R-squared: 0.749, Adjusted R-squared: 0.7434

F-statistic: 134.3 on 2 and 90 DF, p-value: < 2.2e-16.

We have used the GCV criterion to estimate the penalty parameter k. The estimate is $\hat{k}_{opt} = 5.7$. The estimated beta parameters using the ridge regression method with $k = \hat{k}_{opt}$ are given below along with their standard errors. By taking $k = 1$ and using the formula for $\hat{\sigma}^2$ above, we get $\hat{\sigma}^2 = 0.2640458$. We have calculated the t-values of $\hat{\beta}_j(k)$'s, but they are to be taken with a grain of salt since the ridge estimates are biased. These t-values can be used to test if $\beta_j(k) = 0$. Since $\beta_j(k) \neq \beta_j$ (even though $\beta_j(k) \approx \beta_j$ if k is small), we cannot use these t-statistics to test if $\beta_j = 0$.

Coef	Estimate	Std.Error	t value
X_1	-0.2089	0.0918	-2.275
X_2	0.0178	0.0830	0.2146
X_3	0.0290	0.0839	0.3457
X_4	0.5075	0.0749	6.775
X_5	0.0438	0.0879	0.4989
X_6	0.1321	0.0949	1.392

Plot of Y against $\hat{Y} = \hat{\mu}(k)$ shows that the fit is reasonable. However, we should note that there is a problem of unequal variance. This issue can be addressed by taking better transformations, e.g., we should have 1/sqrt(price) as the independent variable instead of logarithm of price. Also there are some nonlinearities. In other words, better transformation of variables and adding nonlinear terms may improve the data analysis.

Mathematical motivations for ridge regression.

In order to motivate the ridge regression method, it may be worthwhile to figure out what may be problematic when $R = X^T X$ is nearly singular. Note that $R = X'X$ is a non negative definite matrix. Using the spectral decomposition theorem we get $R = \sum_{1 \le j \le p} \lambda_j v_j v_j^T$, where $\lambda_1 \ge \cdots \ge \lambda_p \ge 0$ are the eigenvalues of R and $v_1, ..., v_p$ are the corresponding normalized eigenvectors (normalized means $v_j^T v_j = 1$ for all j). If R is nonsingular, then all the eigenvalues are positive.

The least squares estimate of β is $\hat{\beta} = R^{-1}X'Y$. We know that $E(\hat{\beta}) = \beta$ and $Cov(\hat{\beta}) = \sigma^2 R^{-1}$. Consequently,

$$\begin{aligned}
E(||\hat{\beta}||^2) &= ||\beta||^2 + E[||\hat{\beta} - \beta||^2] + 2E[(\beta^T(\hat{\beta} - \beta)] \\
&= ||\beta||^2 + E[tr((\hat{\beta} - \beta)^T(\hat{\beta} - \beta)] \\
&= ||\beta||^2 + tr[Cov(\hat{\beta})] = ||\beta||^2 + tr[\sigma^2 R^{-1}] \\
&= ||\beta||^2 + \sigma^2 tr(R^{-1}) = ||\beta||^2 + \sigma^2 \sum_{1 \le j \le p} 1/\lambda_j
\end{aligned}$$

If R is singular and and has rank $q < p$, then the smallest $p - q$ eigenvalues are equal to zero. In such a case there is no unique least squares solution of the normal equation. Ads a matter of fact the normal equation $X^T X\beta = X^T Y$ has infinite number of solutions and the quantity $\sum_{p-q+1 \le j \le p} 1/\lambda_j = \infty$. Whereas if R is close to singularity, the smallest eigenvalues are quite small and consequently, the value of $\sum_{1 \le j \le p} 1/\lambda_j$ is large. This tells us $E(||\hat{\beta}||^2)$ is large even if $||\beta||^2$ is not large. By Markov inequality we can then show the probability that $||\hat{\beta}||^2$ is large is not negligible. This result provides the motivation to look for estimators of β for which the length $||\beta||$ is not large.

Appendix.

Properties of ridge estimate of β.

One of the first things to note about $\hat{\beta}(k)$ is that it is a biased estimate of β. As a matter of fact the risk $E(||\hat{\beta}(k) - \beta||^2)$ splits into two parts: a part involving the bias and the other involving variance. It turns out that the variance part is smaller than that of the least squares estimate $\hat{\beta}$, which is unbiased. The main idea is that while we introduce some bias when using the ridge estimate, but we gain in variance. It can be shown that there is some $k > 0$ at which the risk of $\hat{\beta}(k)$ is smaller than that of the least squares estimate $\hat{\beta}(0)$. The following results make some of these ideas clear.

Fact 1. Let $\beta(k) = E(\hat{\beta}(k)) = (R + kI)^{-1}R\beta$. Then
(i) $\beta(k) = \beta - k(R + kI)^{-1}\beta$, (ii) $Cov(\hat{\beta}(k)) = \sigma^2(R + kI)^{-1}R(R + kI)^{-1}$.

Fact 2. Let $SSE(k) = ||Y - X\hat{\beta}(k)||^2$, $H(k) = X(R + kI)^{-1}X^T$ and $\mu(k) = X\beta(k)$. Then

$$E(SSE(k)) = \sigma^2 tr((I - H(k))^2) + ||\mu - \mu(k)||^2.$$

Let us now analyze the risk of the estimate of β a bit further. The risk can be decomposed into two parts: square of bias and variance. Note that

$$\begin{aligned}
E(||\hat{\beta}(k) - \beta||^2) &= E(||\hat{\beta}(k) - \beta(k) + \beta(k) - \beta||^2) \\
&= E(||\hat{\beta}(k) - \beta(k)||^2) + ||\beta(k) - \beta||^2 + 2E((\hat{\beta}(k) - \beta(k))'(\beta(k) - \beta))
\end{aligned}$$

Clearly, the cross product term $E((\hat{\beta}(k)-\beta(k))'(\beta(k)-\beta))$ equals zero. Hence

$$\begin{aligned} E(||\hat{\beta}(k)-\beta(k)||^2) &= E(||\hat{\beta}(k)-\beta(k)||^2) + ||\beta(k)-\beta||^2 \\ &= \phi_2(k) + \phi_1(k) := variance + bias^2. \end{aligned}$$

Now using the spectral decomposition of R we get

$$\begin{aligned} E(||\hat{\beta}(k)-\beta(k)||^2) &= tr(\sigma^2(R+kI)^{-1}R(R+kI)^{-1}) \\ &= \sigma^2 tr((R+kI)^{-2}R) = \sigma^2 \sum \lambda_j/(k+\lambda_j)^2 = \phi_2(k), \text{ say.} \end{aligned}$$

Now using part (i) of Fact 1 we get

$$||\beta(k)-\beta||^2 = ||-k(R+kI)^{-1}\beta||^2 = k^2 \sum (v_j^T\beta)^2/(k+\lambda_j)^2 = \phi_1(k), \text{ say.}$$

Hence we have

$$E(||\hat{\beta}(k)-\beta||^2) = \phi_2(k) + \phi_1(k),$$

where $\phi_2(k)$ corresponds to the variance part and $\phi_1(k)$ corresponds to the bias part. In order to understand ridge regression better, we will write down a few properties of $\phi_1(k)$ and $\phi_2(k)$. These properties basically say that the variance (i.e., ϕ_2) decreases as $k \uparrow$, whereas the bias^2 (i.e., ϕ_1) increases as $k \uparrow$. There exists $k > 0$ such that $E(||\hat{\beta}(k)-\beta||^2) < E(||\hat{\beta}(0)-\beta||^2)$, i.e., it is possible to improve over the ordinary least squares method using ridge regression.

Properties of ϕ_1 and ϕ_2.

1) For any $k > 0$ we have,

$$||\beta||^2 = \phi_1(\infty) > \phi_1(k) > \phi_1(0) = 0,$$

and

$$0 = \phi_2(\infty) < \phi_2(k) < \phi_2(0) = \sigma^2 \sum 1/\lambda_j.$$

2) The derivative of ϕ_1 with respect to k is

$$\phi_1'(k) = 2k \sum \lambda_j (v_j^T\beta)^2/(k+\lambda_j)^3.$$

Note that $\phi_1'(k) > 0$ for any $k > 0$ which implies that ϕ_1 is an increasing function k.

3) It is not difficult to see that

$$\phi_2'(k) = -2\sigma^2 \sum \lambda_j/(k+\lambda_j)^3, \quad \phi_2''(k) = 6\sigma^2 \sum \lambda_j/(k+\lambda_j)^4.$$

Since $\phi_2'(k) < 0$ for all $k \geq 0$, we can conclude that ϕ_2 is a decreasing function. Since $\phi_2''(k) > 0$ for all $k \geq 0$, we can conclude that ϕ_2 is a convex function.

4) Since $\phi_1'(k) \approx 0$ and $\phi_2'(k) < 0$ when $k \approx 0$, calculus tells us

$$\begin{gathered} \phi_2(k) + \phi_1(k) < \phi_2(0) + \phi_1(0), \text{i.e.,} \\ E(||\hat{\beta}(k)-\beta||^2) < E(||\hat{\beta}(0)-\beta||^2), \end{gathered}$$

when $k \approx 0$.

Mean square error

Let us now examine how the ridge regression method performs in estimating the mean $\mu = X\beta$. We will use the expected squared error loss to judge this performance. Let

$$\begin{aligned} L(k) &= E[||X\hat{\beta}(k) - X\beta||^2] = E[||X\hat{\beta}(k) - X\beta(k)||^2] + ||X\beta(k) - X\beta||^2] \\ &= \sigma^2 trace((H(k)^2) + ||\mu(k) - \mu||^2 := variance + bias^2 \\ &= \sigma^2 \sum \lambda_j^2/(k+\lambda_j)^2 + k^2 \sum (v_j^T \beta)^2 \lambda_j/(k+\lambda_j)^2 \\ &= L_1(k) + L_2(k), \text{ say.} \end{aligned}$$

Under some conditions it can be shown that $L(k)$ is a convex function k and it is bathtub shaped. When $k = 0$ we get the ordinary least squares (OLS) estimate of β. Note that $L_1(k)$ is the variance term and $L_2(k)$ can be described as the bias-squared term. A not so difficult calculation will show that the derivatives of L_1 and L_2 (differentiating with respect to k)

$$L_1'(k) = -2\sigma^2 \sum \lambda_j^2/(k+\lambda_j)^3, \; L_2'(k) = 2k \sum (v_j^T\beta)^2 \lambda_j/(k+\lambda_j)^3.$$

These expressions for the derivatives of L_1 and L_2 clearly tell us that the variance term $L_1(k)$ decreases as k increases, whereas the bias term $L_2(k)$ increases as k increases. Note that

$$L'(0) = -2\sigma^2 \sum 1/\lambda_j < 0.$$

This mean that for some $k > 0$ we have $L(k) < L(0) = p\sigma^2$. This in turn implies that there a ridge estimator for the mean vector $\mu = X\beta$ for some $k > 0$ which is superior to the ordinary least squares estimate.

Estimation of the penalty parameter k

How do we estimate the penalty parameter k? We need to do it using the data that is available. There are a number of methods but we will outline only one here. We have already seen that

$$L(k) = E[||X\hat{\beta}(k) - X\beta||^2] = \sigma^2 tr(H(k)^2) + ||\mu(k) - \mu||^2$$

Ideally we would like to find the value $k = k^*$ at which $L(k)$ is minimized and then declare k^* to be our choice of the penalty parameter k. However, $L(k)$ involves σ^2 and β which are unknown. So we can then try to estimate $L(k)$ by $\hat{L}(k)$ and then minimize $\hat{L}(k)$ over k. If the minimum is achieved at $k = \hat{k}^*$, then we can take $\hat{k}^*$ to be our estimate of the optimal value of penalty parameter k^*.

Now consider the residual sum of squares $SSE(k)$ when the mean vector $\mu = X\beta$ is estimated by $X\hat{\beta}(k)$. Note that

$$E(SSE(k)) = E[||Y - X\hat{\beta}(k)||^2] = n\sigma^2 + L(k) - 2\sigma^2 tr(H(k))$$

If σ were known an estimate of $L(k)$ would be

$$\widetilde{L}(k) = SSE(k) + 2\sigma^2 tr(H(k)) - n\sigma^2.$$

Since the last term does not involve k, it plays no role in the minimization (over k) and hence, if we have a "good" estimate $\hat{\sigma}^2$ of σ^2, an estimate of $L(k)$ would be

$$\hat{L}(k) = SSE(k) + 2\hat{\sigma}^2 tr(H(k)) - n\sigma^2$$

Ignoring the last term we then arrive at a criterion function

$$C(k) = SSE(k) + 2\hat{\sigma}^2 tr(H(k))$$

which is to be minimized over k in order to get an estimate of the optimal penalty parameter k^*.

There are many different types of "good" estimators of σ^2 for different types of models. However, we will concentrate on only two of the well known methods.

Akaike-type criterion:

Noting that

$$E(SSE(k)) = \sigma^2 tr[(I - H(k))^2] + ||\mu(k) - \mu||^2,$$

a reasonable estimate of σ^2 is given by

$$\hat{\sigma}^2(k) = SSE(k)/tr[(I - H(k))^2]$$

provided that $||\mu(k) - \mu||^2$ is small which which is true if k small.

This leads to Akaike type criterion which is

$$C_{AKA}(k) = SSE(k) + 2\hat{\sigma}^2(k)tr(H(k)) = SSE(k) + 2\hat{\sigma}^2(k)tr(H(k))$$

Mallows' type criterion:

Mallows type criterion typically does not specify how σ^2 is to be chosen. However, the following method can be used. Let k_0 be reasonably small value of k so that

$$\hat{\sigma}^2(k_0) = SSE(k_0)/tr[(I - H(k_0))^2]$$

is almost an unbiased estimator of σ^2. Then a Mallows type criterion is

$$C_{MAL}(k) = SSE(k) + 2\hat{\sigma}^2(k_0)tr(H(k))$$

Cross-validation and generalized cross-validation.

Before we discuss tis method, let us first provide the intuitive background. Suppose that we have n iid observation (Y_i, x_i), $i = 1, \ldots, n$, where x_i's are p dimensional (they form the rows of the matrix X in the linear model), on the basis of which a regression model has been estimated and suppose $\hat{\beta}$ is the estimate of β. Now we may want to see if the regression model that has been developed on the basis of the data predicts predicts a future observation (Y_{n+1}, x_{n+1}) well, where it is understood that (Y_{n+1}, x_{n+1}) is independent of the data but has the sample distribution as (Y_i, x_i), $i = 1, \ldots, n$, i.e., we want to see if Y_{n+1} is well predicted by $x_{n+1}^T\hat{\beta}$. The prediction error is defined to be

$$PE = E[(Y_{n+1} - x_{n+1}^T\hat{\beta})^2].$$

The method of cross-validation tries to estimate this prediction error PE.

Cross-validation (or leave-one cross-validation) involves leaving out one of the $n = 93$ cases out (say the i^{th} case), calculating $\hat{\beta}^{(-i)}(k)$ on the basis of the $n-1$ observations and then use this estimate $\hat{\beta}^{(-i)}(k)$ to predict the i^{th} case. More formally, denote the i^{th} row of X by x_i^T. For a given penalty parameter k, we obtain the ridge estimate $\hat{\beta}^{(-i)}(k)$ on the basis of all the observations except the i^{th} case (Y_i, x_i). Then we predict Y_i using $x_i^T\hat{\beta}^{(-i)}(k)$. So an estimate of the prediction error is thus $(Y_i - x_i^T\hat{\beta}^{(-i)}(k))^2$. If we repeat this for $i = 1, \ldots, n$, we will end up with n estimates of the prediction error. So the cross-validated estimate of the prediction error is thus

$$PE^{CV}(k) = n^{-1}\sum_{i=1}^{n}(Y_i - x_i^T\hat{\beta}^{(-i)}(k))^2.$$

We can then minimize $PE^{CV}(k)$ over k and find the value $\hat{k}_{opt}$ at which $PE^{CV}(k)$ attains its minimum. We can then use $\hat{k}_{opt}$ as the penalty parameter for our ridge regression.

It looks as if one needs to calculate, for each $k > 0$, the ridge estimate of β n times. Fortunately, there is a nice clear formula, which reduces the need for much calculation. If we denote $X^{(-i)}$ as the $(n-1) \times p$ matrix which is obtained by deleting the i^{th} row of X and $Y^{(-i)}$ as the $n-1$-dim vector which is obtained by deleting the i^{th} row of Y, then the $\hat{\beta}^{(-i)}(k)$ is a solution of

$$(X^{(-i)T}X^{(-i)} + kI)\beta = X^{(-i)T}Y^{(-i)}.$$

Now note that

$$X^{(-i)T}X^{(-i)} = X^TX - x_ix_i^T,\; X^{(-i)T}Y^{(-i)} = X^TY - x_iY_i.$$

Thus we can recast the above formula as

$$\begin{aligned}&(X^TX - x_ix_i^T + kI)\beta = X^TY - x_iY_i, \text{ or}\\ &(X^TX + kI - x_ix_i^T)\beta = X^TY - x_iY_i.\end{aligned}$$

We need to invert the matrix $X^TX + kI - x_ix_i^T$. There is a famous formula for updating the inverse of rank-one update of a matrix. Let B be a $m \times m$ nonsingular matrix and a be a $m \times 1$ vector. Then, assuming that $a^TBa \neq 1$ and B is invertible, we have

$$(B - aa^T)^{-1} = B^{-1} + (1 - a^TB^{-1}a)^{-1}B^{-1}aa^TB^{-1}.$$

Apply this result with $B = X^TX + kI$ and $a = x_i$ to get

$$\begin{aligned}\hat{\beta}^{(-i)}(k) &= (X^TX + kI - x_ix_i^T)^{-1}(X^TY - x_iY_i)\\ &= [B^{-1} + (1 - x_i^TBx_i)^{-1}B^{-1}x_ix_i^TB^{-1}](X^TY - x_iY_i)\\ &= [I + (1 - x_i^TBx_i)^{-1}B^{-1}x_ix_i^T][\hat{\beta}(k) - B^{-1}x_iY_i].\end{aligned}$$

Thus denoting $x_i^T B x_i = x_i^T (R + kI)^{-1} x_i$ by h_{ii} we get

$$\begin{aligned}
x_i^T \hat{\beta}^{(-i)}(k) &= [x_i^T + ((1 - x_i^T B x_i)^{-1} x_i^T B^{-1} x_i x_i^T](\hat{\beta}(k) - B^{-1} x_i Y_i) \\
&= [x_i^T + (1 - h_{ii})^{-1} h_{ii} x_i^T](\hat{\beta}(k) - B^{-1} x_i Y_i) \\
&= [1 + (1 - h_{ii})^{-1} h_{ii}](x_i^T \hat{\beta}(k) - h_{ii} Y_i) \\
&= (1 - h_{ii})^{-1}(\hat{\mu}_i(k) - h_{ii} Y_i), \text{ and hence} \\
Y_i - x_i^T \hat{\beta}^{(-i)}(k) &= Y_i - (1 - h_{ii})^{-1}(\hat{\mu}_i(k) - h_{ii} Y_i) = (1 - h_{ii})^{-1}(Y_i - \hat{\mu}_i(k)).
\end{aligned}$$

Thus

$$PE^{CV}(k) = n^{-1} \sum (Y_i - x_i^T \hat{\beta}^{(-i)}(k))^2 = n^{-1} \sum (1 - h_{ii})^{-2} (Y_i - \hat{\mu}_i(k))^2.$$

This shows that one does not need to calculate $\hat{\beta}^{(-i)}(k)$ for each i in order to obtain $PE^{CV}(k)$.

Now if one replaces h_{ii}'s by their average, then a modified version of PE^{CV} is obtained and it is called generalized cross-validation estimate of the prediction error. Note that h_{ii} is the i^{th} diagonal element of $X(R + kI)^{-1} X^T = H(k)$. Thus the average of h_{ii}'s equal $\bar{h} = tr(H(k))/n$. Thus we arrive at the GCV formula

$$\begin{aligned}
PE^{GCV}(k) &= n^{-1} \sum (1 - \bar{h})^{-2} (Y_i - \hat{\mu}_i(k))^2 = n^{-1} \sum (Y_i - \hat{\mu}_i(k))^2 (1 - \bar{h})^{-2} \\
&= n^{-1} SSE(k)(1 - \bar{h})^{-2} = n^{-1} SSE(k)/[1 - tr(H(k))]^2
\end{aligned}$$

The GCV criterion presented before is simply a rescaled version of the criterion derived here.

How good is $\hat{k}^*$?

There are mathematical results which prove that the criteria described above for estimating k^* work quite well. Here is a mathematical result which hold under appropriate technical assumptions.

Fact3: Denote by $\hat{k}^*$ the estimate of k^* which is obtained by using any of the four criteria described above. Then under appropriate technical conditions it can be shown that the ratio

$$||X\hat{\beta}(\hat{k}^*) - X\beta||^2 / [\min_{k \geq 0} ||X\hat{\beta}(k) - X\beta||^2]$$

converges in probability to 1.

Technical Notes:

Expression for $L(k)$:

$$\begin{aligned}
L(k) = E[||X\hat{\beta}(k) - X\beta||^2] &= E[||X\hat{\beta}(k) - X\beta(k) + X\beta(k) - X\beta||^2 \\
&= E[||X\hat{\beta}(k) - X\beta(k)||^2] + ||X\beta(k) - X\beta||^2 \\
&= E[||X(R + kI)^{-1} X^T \varepsilon||^2] + ||\mu(k) - \mu||^2 \\
&= E[||H(k)\varepsilon||^2] + ||\mu(k) - \mu||^2 = \sigma^2 tr(H(k)^2) + ||\mu(k) - \mu||^2.
\end{aligned}$$

Expression for $E(SSE(k))$:

$$\begin{aligned}
E(SSE(k)) &= E[||Y - X\hat{\beta}(k)||^2 \\
&= E[||\varepsilon - (X\hat{\beta}(k) - X\beta)||^2] \\
&= E(||\varepsilon||^2) + E[||X\hat{\beta}(k) - X\beta||^2] - 2E\varepsilon^T(X\hat{\beta}(k) - X\beta) \\
&= n\sigma^2 + L(k) - 2E(\varepsilon^T(X\hat{\beta}(k) - X\beta(k))) - 2E(\varepsilon^T(X\beta(k) - X\beta)) \\
&= n\sigma^2 + L(k) - 2E(\varepsilon^T X(R + kI)^{-1}X^T\varepsilon) - 0 \\
&= n\sigma^2 + L(k) - 2E(\varepsilon^T H(k)\varepsilon) = n\sigma^2 + L(k) - 2\sigma^2 tr(H(k))
\end{aligned}$$

A more detailed expression for $E(SSE(k))$:

Substituting the expression for $L(k)$ in that of $E(SSE(k))$ we get

$$\begin{aligned}
E(SSE(k)) &= n\sigma^2 + L(k) - 2\sigma^2 tr(H(k)) \\
&= n\sigma^2 + \sigma^2 tr(H(k)^2) + ||\mu(k) - \mu||^2 - 2\sigma^2 tr(H(k)) \\
&= \sigma^2[n + tr(H(k)^2) - tr(H(k))] + ||\mu(k) - \mu||^2 \\
&= \sigma^2 tr[(I - H(k))^2] + ||\mu(k) - \mu||^2.
\end{aligned}$$

Figure 1: Matrix plot of car 93: untransformed

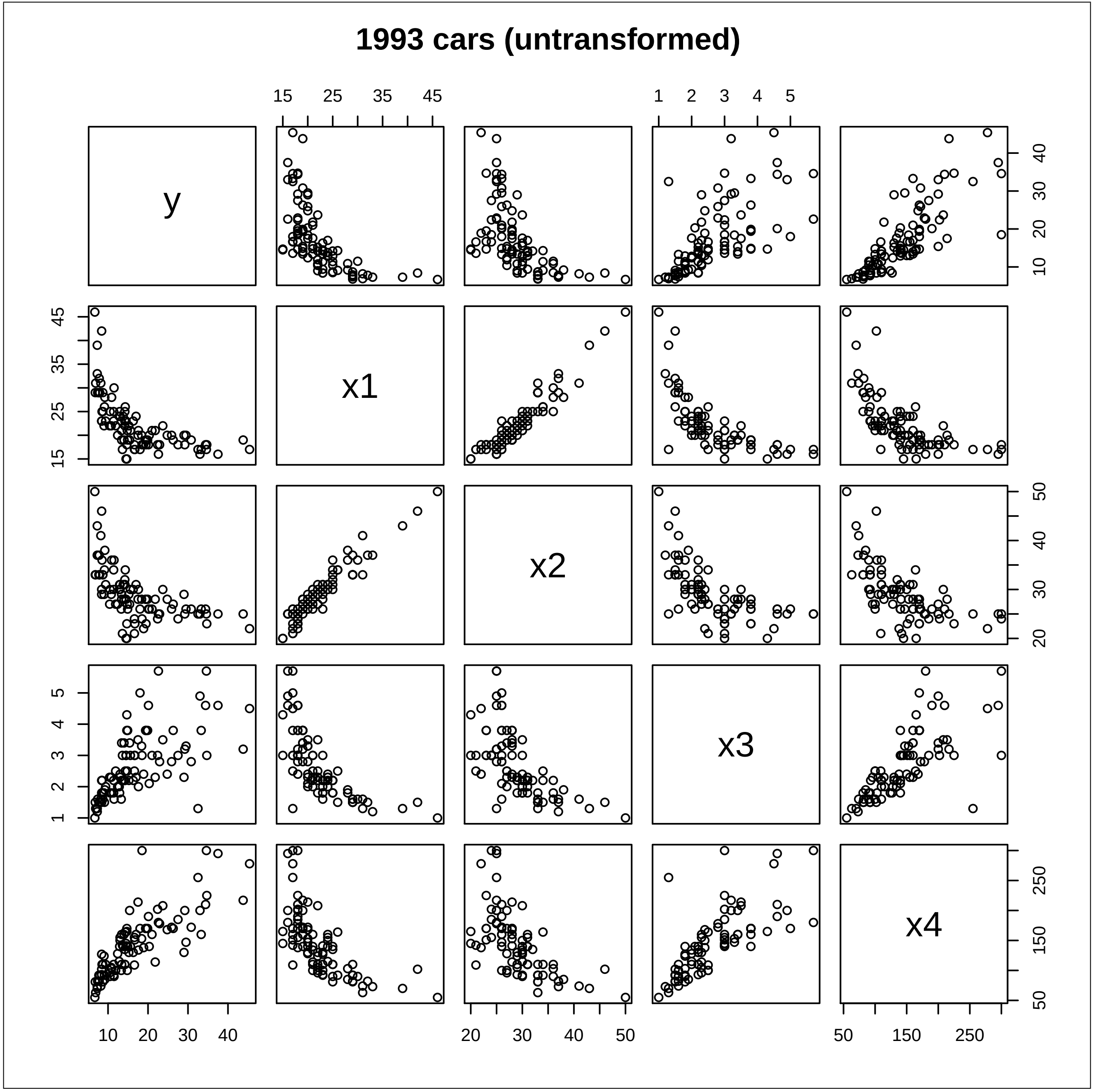


Figure 2: Matrix plot of car 93: transformed

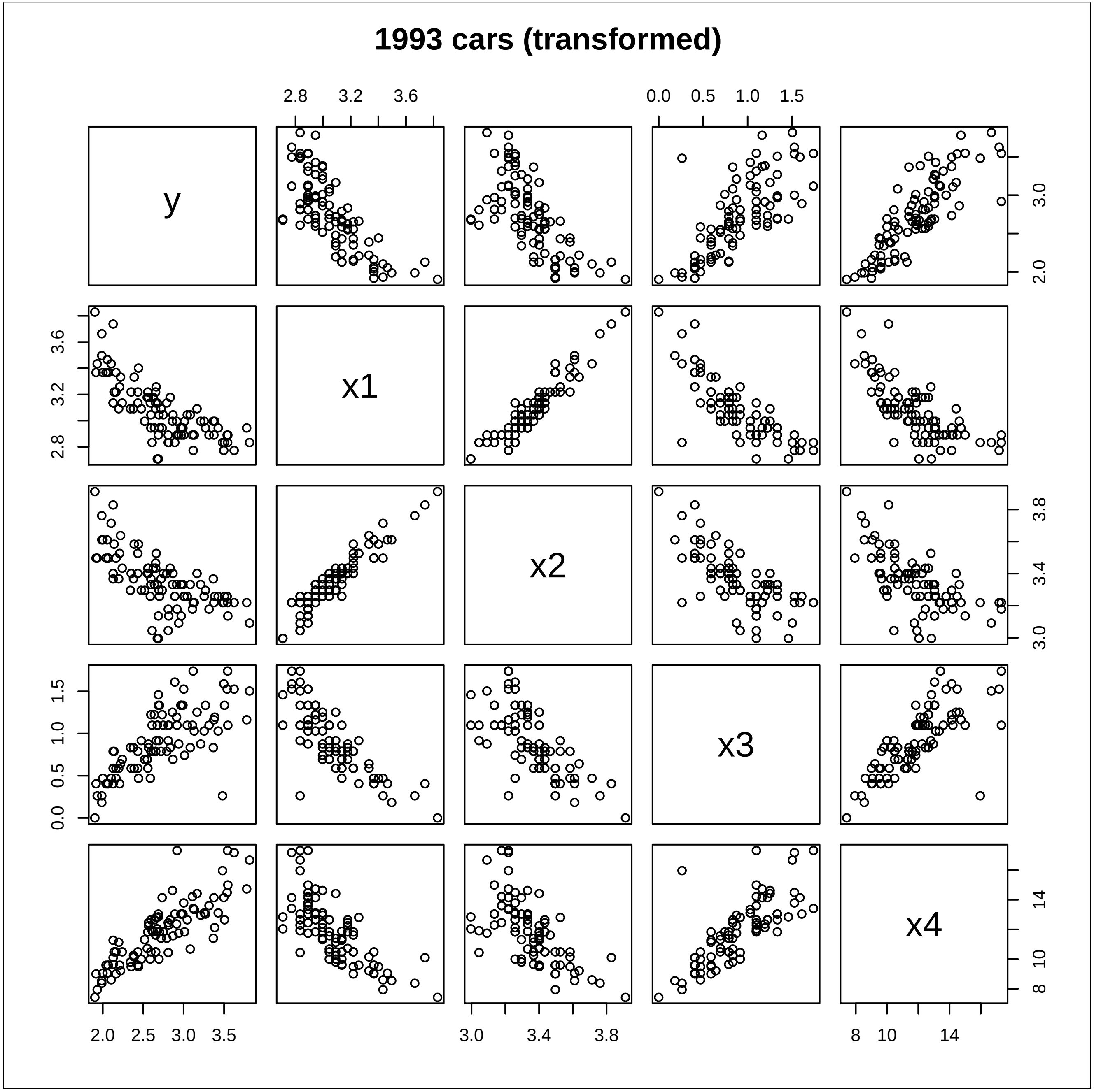
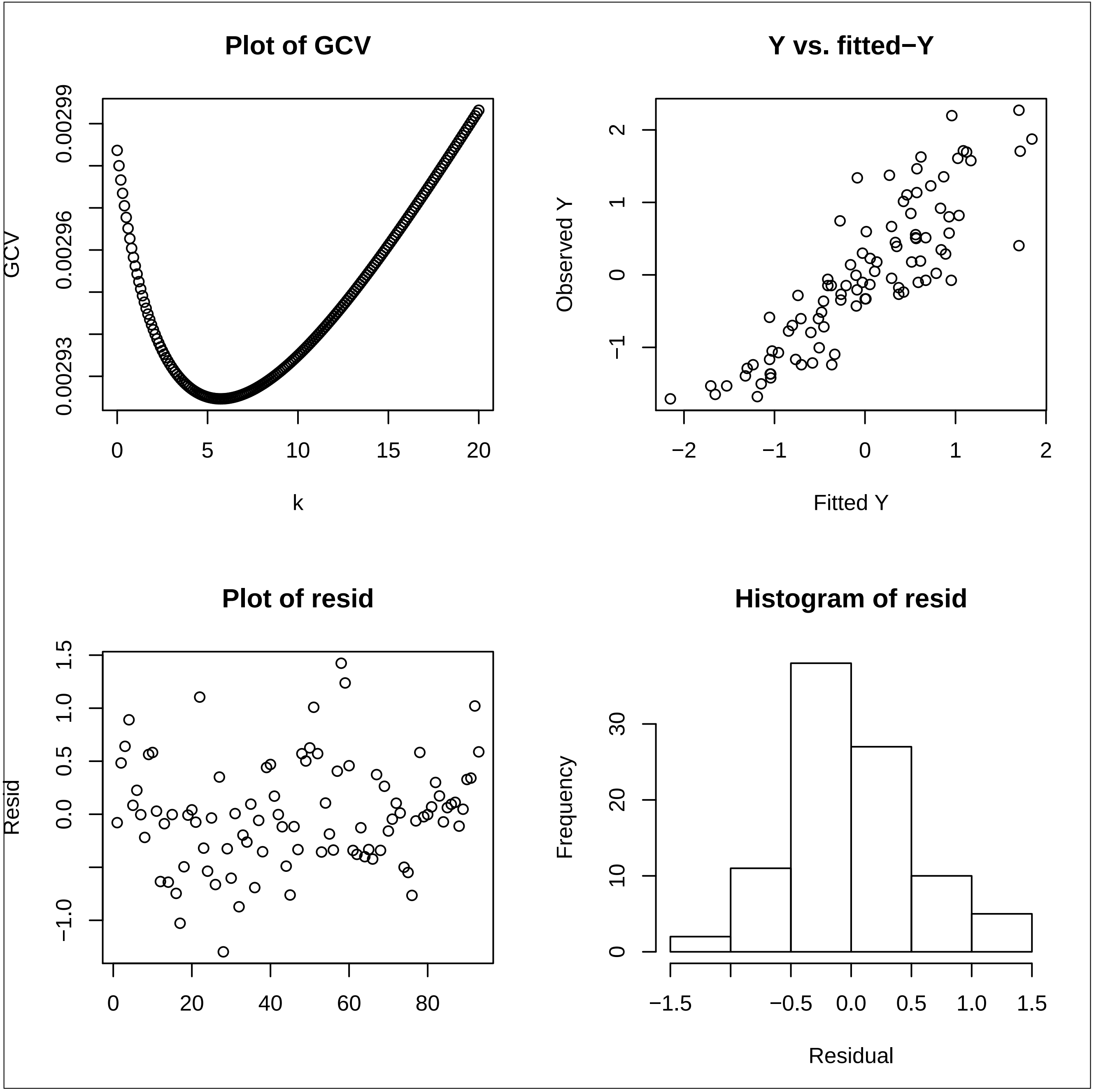


Figure 3: Ridge Regression Analysis of cars93


Plot of GCV
GCV
k
Y vs. fitted−Y
Observed Y
Fitted Y
Plot of resid
Resid
Histogram of resid
Frequency
Residual