Chapter 11 Remedial Measures

- Transformation
- Weighted least squares
- Ridge regression
- Lasso
- Regression Trees (Random Forests)
- Bootstrapping

Weighted least squares

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i$$

$\beta_0, \beta_1, \cdots, \beta_{p-1}$ are parameters

$x_{i1}, x_{i2}, \cdots, x_{i,p-1}$ are known constants

ϵ_i are independent $N(0, \sigma_i^2)$

$i = 1, \cdots, n$

- Unequal Variance

$$\begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

- Least square doesn’t work, what should we do?

Use maximum likelihood

$$
\begin{aligned}
L(\boldsymbol{\beta}) &= \prod_{i=1}^{n} \frac{1}{(2\pi\sigma_i^2)^{1/2}} \text{exp}[-\frac{1}{2\sigma_i^2}(Y_i - \beta_0 - \beta_1 X_{i1} \\
&- \cdots - \beta_{p-1}X_{i,p-1})^2]
\end{aligned}
$$

Define the ith weight to be

$$w_i = \frac{1}{\sigma_i^2}$$

The likelihood is

$$
\begin{aligned}
L(\boldsymbol{\beta}) &= \left[\prod_{i=1}^{n} \left(\frac{w_i}{2\pi}\right)^{1/2}\right] \text{exp}[-\frac{1}{2}\sum_{i=1}^{n} w_i(Y_i - \beta_0 - \beta_1 X_{i1} \\
&- \cdots - \beta_{p-1}X_{i,p-1})^2]
\end{aligned}
$$

Suppose that σ_i^2's are known, the log likelihood is a constant plus

$$Q_w = \sum_{i=1}^{n} w_i (Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_{p-1} X_{i,p-1})^2$$

Criterion is same as least squares, except each squared residual is weighted by w_i—hence the weighted least squares criterion.

The coefficient vector b_w that minimizes Q_w is the vector of weighted least squares estimates

Let

$$\mathbf{W} = \begin{bmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \vdots & w_n \end{bmatrix}$$

- The regression coefficients with weights are

$$\mathbf{b}_w = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{W}\mathbf{Y})$$

$$\sigma^2(\mathbf{b}_w) = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}$$

- Least squares minimizes the sum of the squared residuals. Weighted least squares minimizes the sum of the squared residuals each multiplied by an appropriate weight.

How about σ_i^2's are unknown?

- If the error variances σ_i^2's are known, the weights are $w_i = 1/\sigma_i^2$
- Otherwise, the variances need to be estimated

  —$\sigma_i^2 = E(\epsilon_i^2) - (E(\epsilon_i))^2 = E(\epsilon_i^2)$

  —The squared residual e_i^2 is an estimator of σ_i^2.

  —The absolute residual $\left|e_i\right|$ is an estimator of the standard deviation σ_i

We can therefore

- Estimate the variance function describing the relation of σ_i^2 to relevant predictor variables: First fitting the regression model using unweighted least squares and then regressing the squared residuals e_i^2 against the appropriate predictor variables

- Estimate the standard deviation function describing the relation of σ_i to relevant predictor variables: First fitting the regression model using unweighted least squares and then regressing the absolute residuals $|e_i|$ against the appropriate predictor variables

- A residual plot against X_1 exhibits a megaphone shape. Regress the absolute residuals against X_1.

- A residual plot against $\hat{Y}$ exhibits a megaphone shape. Regress the absolute residuals against $\hat{Y}$.

- A plot of the squared residuals against X_3 exhibits an upward tendency. Regress the squared residuals against X_3.

- A plot of the squared residuals against X_2 suggests variance increases rapidly with increase in X_2 up to a point and then increases more slowly. Regress the absolute residuals against X_2 and X_2^2.

- After the variance function or the standard deviation function is estimated,

  the fitted value from this function are used to obtain the estimated weights:

  —$w_i = \dfrac{1}{\hat{v}_i}$, where $\hat{v}_i$ is fitted value from variance function

  —$w_i = \dfrac{1}{\hat{s}_i^2}$, where $\hat{s}_i$ is fitted value from standard deviation function

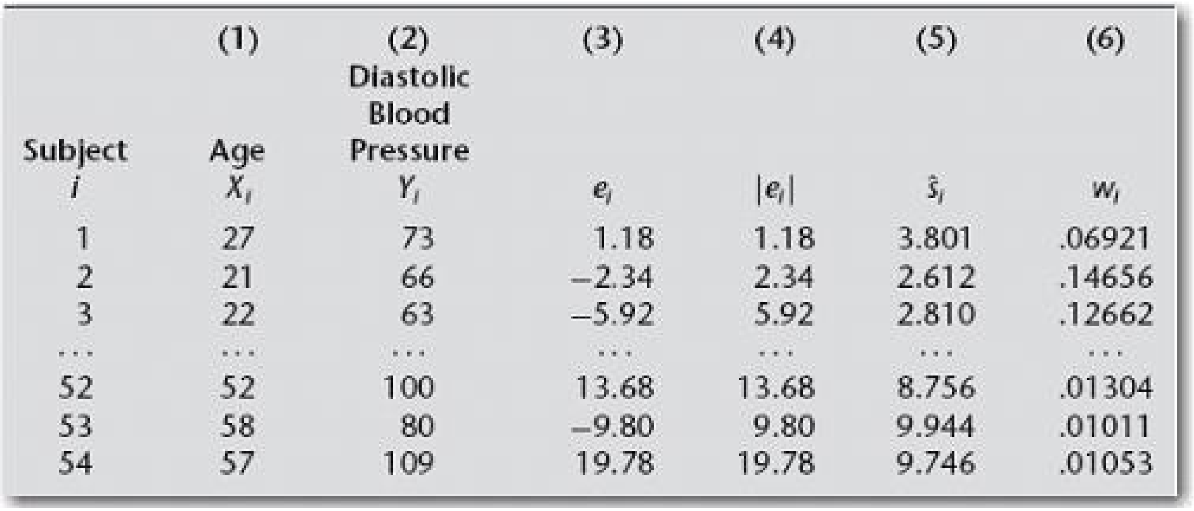
Summary

- Fit the regression model by unweighted least squares and analyze the residuals
- Estimate the variance function or the standard deviation function by regressing either the squared residuals or the absolute residuals on the appropriate predictor(s)
- Use the fitted values from the estimated variance or standard deviation function to obtain the wights w_i
- Estimate the regression coefficients using these weights

Example: page 427

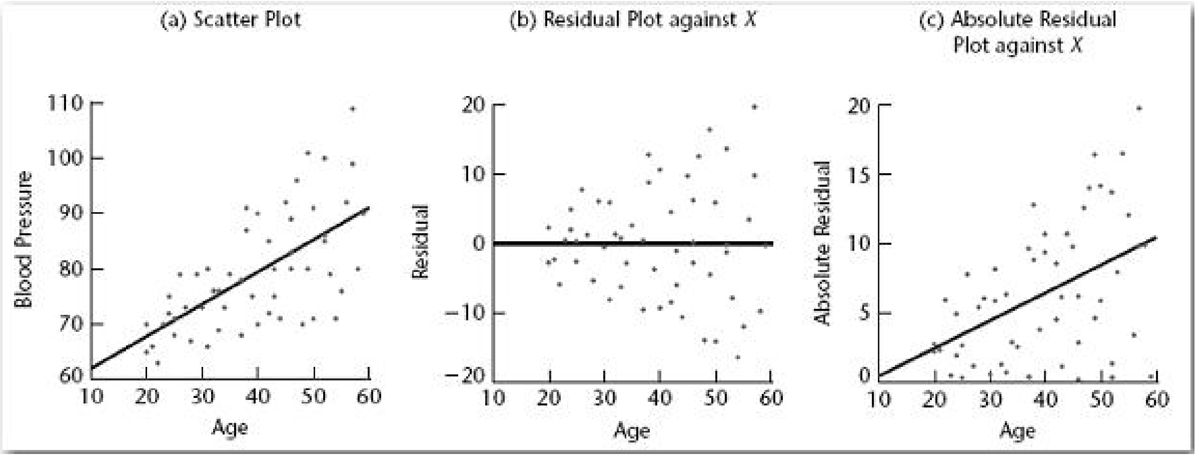
A health researcher, interested in studying the relationship between diastolic blood pressure and age among healthy adult women 20 to 60 years old, collected data on 54 subjects. Portion of the data are as follows

Figure 1: Example



Subject	(1) Age	(2) Diastolic Blood Pressure	(3)	(4)	(5)	(6)
i	X_i	Y_i	e_i	$\|e_i\|$	$\hat{s}_i$	w_i
1	27	73	1.18	1.18	3.801	.06921
2	21	66	−2.34	2.34	2.612	.14656
3	22	63	−5.92	5.92	2.810	.12662
...	...	...	...	...	...	...
52	52	100	13.68	13.68	8.756	.01304
53	58	80	−9.80	9.80	9.944	.01011
54	57	109	19.78	19.78	9.746	.01053

Figure 2: Scatterplot, residual plot and absolute residual plots, example page 472



- Scatter plot of the data suggests a linear relationship between diastolic blood pressure and age but also indicates that the error term variance increases with age.

- Use unweighted regression

$$\hat{Y} = 56.157 + 0.58003X$$

- Figure (c) suggests that a linear relation between the error standard deviation and X may be reasonable.

$$\hat{s}_i = -1.54946 + 0.198172X_i$$

- For case 1, $X_1 = 27$, the fitted value is

$$\hat{s}_1 = -1.54946 + 1.98172 * (27) = 3.801$$

$$w_1 = \frac{1}{\hat{s}_1^2} = \frac{1}{3.801^2} = 0.0692$$

- The weighted estimated regression function is

$$\hat{Y} = 55.566 + 0.59634X$$

95% CI for β_1 is

$$0.437 \leq \beta_1 \leq 0.755$$

Drawbacks and Advantages

- WLS estimates are minimum variance, unbiased.
- If you use Ordinary Least Squares (OLS) when variance is not constant, estimates are still unbiased, just not minimum variance.
- If you have replicates at each unique X category, you can just use the sample standard deviation of the responses at each category to determine the weight for any response in the category.
- R_2 has no clear cut meaning here.
- Must use the standard deviation function value (instead of s) for confidence intervals for prediction

Ridge regression

- Remedy multicollinearity problems
- Modifying the method of least squares to allow biased estimators of the regression coefficients
- Small bias but more precise than an unbiased estimator

  —Shrinkage estimation: Reduce the variance of the parameters by shrinking them (a bit) in absolute magnitude. This will introduce some bias, but may reduce the MSE overall.

  —Recall: MSE = bias squared plus variance:

$$E\{\hat{Y}_i - \mu_i\}^2 = (E\{\hat{Y}_i\} - \mu_i)^2 + V(\hat{Y}_i)$$

Ridge Estimators

- Normal equation for ordinary least squares

$$(\mathbf{X}'\mathbf{X})\mathbf{b} = \mathbf{X}'\mathbf{Y}$$

- Correlation transformation

$$y_i^* = \frac{1}{\sqrt{n-1}}\left(\frac{y_i - \bar{y}}{s_y}\right)$$

$$x_{ik}^* = \frac{1}{\sqrt{n-1}}\left(\frac{x_{ik} - \bar{x}_k}{s_k}\right)(k = 1, \cdots, p-1)$$

- The regression model with the transformed variable y^* and x_k^* as defined by the correlation transformation is called a standardized regression model

$$y_i^* = \beta_1^* x_{i1}^* + \cdots + \beta_{p-1}^* x_{i,p-1}^* + \epsilon_i^*$$

- Least square normal equations are

$$r_{\mathbf{XX}}\mathbf{b} = r_{\mathbf{YX}}$$

$$r_{\mathbf{XX}} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1,p-1} \\ r_{21} & 1 & \cdots & r_{2,p-1} \\ \vdots & \vdots & & \vdots \\ r_{p-1,1} & r_{p-1,2} & \cdots & 1 \end{bmatrix}$$

$$r_{\mathbf{YX}} = \begin{bmatrix} r_{\mathbf{Y}1} \\ r_{\mathbf{Y}2} \\ \vdots \\ r_{\mathbf{Y}p-1} \end{bmatrix}$$

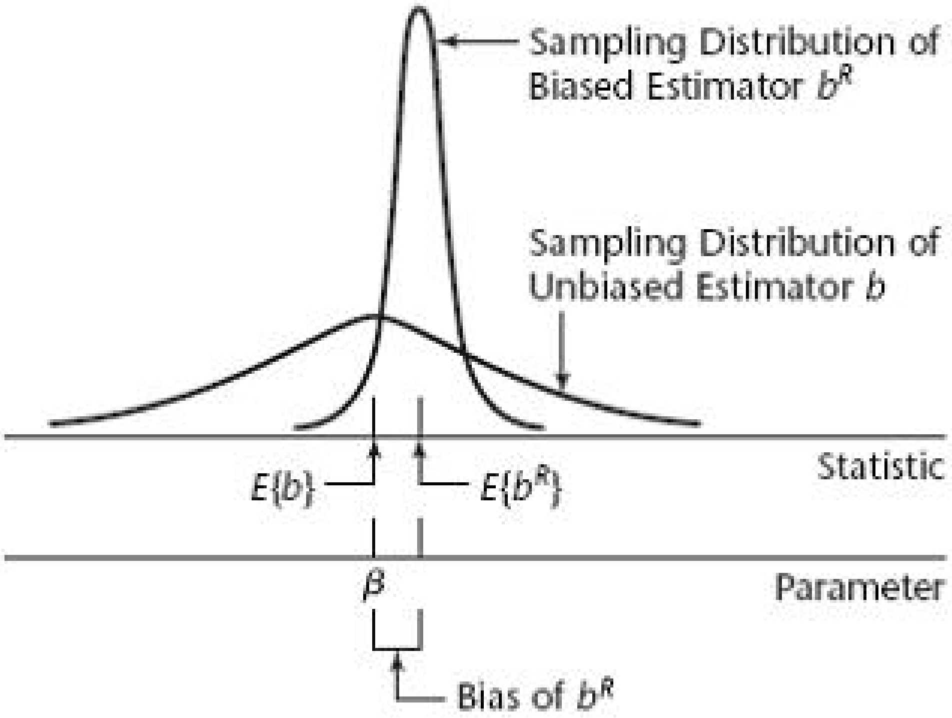
- Ridge normal equation

$$(r_{\mathbf{XX}} + c\mathbf{I})\mathbf{b}^R = r_{\mathbf{YX}}$$

  $\mathbf{b}^R$ is the vector of the standardized ridge regression coefficients b_k^R

- $$\mathbf{b}^R = (\mathbf{r}_{\mathbf{XX}} + c\mathbf{I})^{-1}\mathbf{r}_{\mathbf{YX}}$$

- c reflects the amount of bias in the estimators

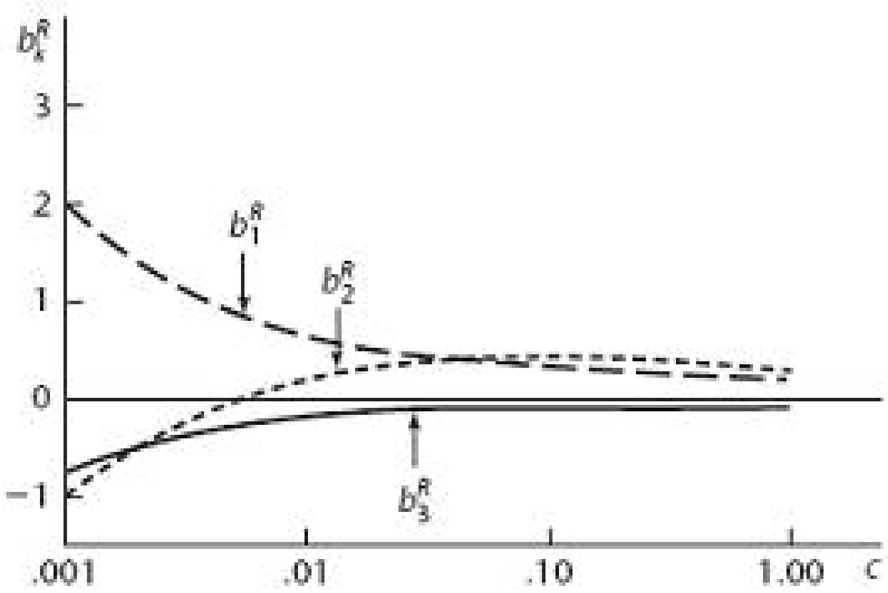


$$E\{b^R - \beta\}^2 = (E\{\hat{b}^R\} - \beta)^2 + V(b^R)$$

Ridge trace, determining the constant c

- Simultaneous plot of the values of the $p-1$ estimated ridge standardized regression coefficients for different values of c, usually between 0 and 1.
  —The estimated regression coefficients b_k^R may fluctuate widely as c is changed slightly from 0, and some may even change signs. Gradually, these wide fluctuations cease and the magnitudes of the regression coefficients tend to move slowly toward zero as c increased further.

Figure 4: Ridge Trace of estimated standardized regression coefficients–bodyfat example with three predictor variables



- $(VIF)_k$ tends to fall rapidly as c is changed from 0, and gradually the $(VIF)_k$ values also tend to change only moderately as c increased further.

- Examine the ridge trace and the VIF values and choose the smallest value of c where it deemed that the regression coefficients first become stable in the ridge trace and the VIF values have become sufficiently small

Example: Bodyfat example with three predictor variables: triceps skinfold thickness, thigh circumference, midarm circumference.

$$\hat{Y} = 117.08 + 4.334X_1 - 2.857X_2 - 2.186X_3$$

b_2 = -2.857 is negative, even though it was expected that amount of body fat is positively related to thigh circumference.

The ridge standardized regression coefficients for selected values of c and VIF are given in the following table,

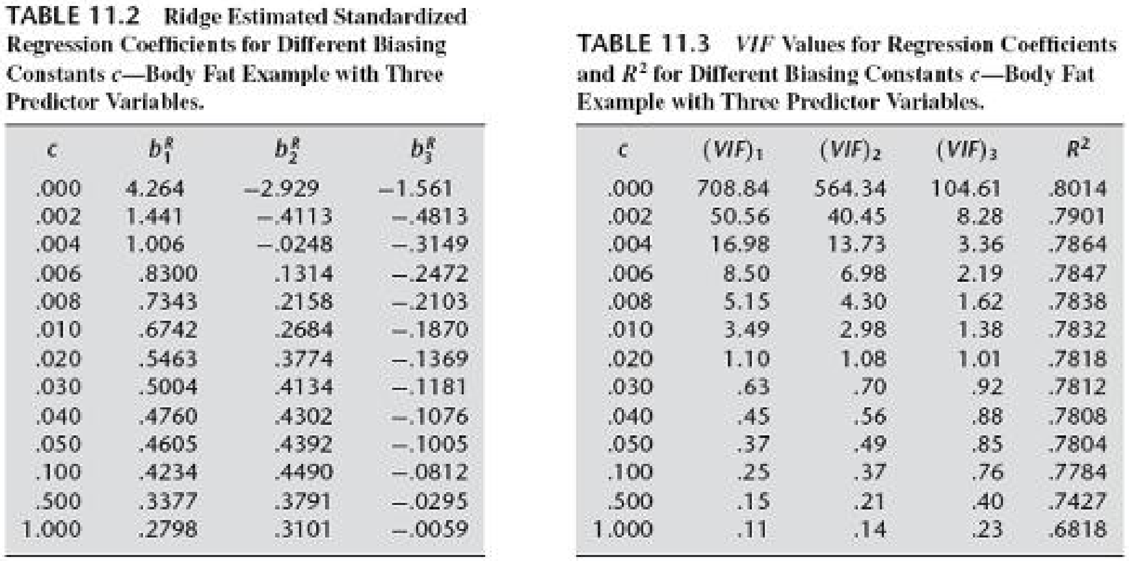


TABLE 11.2 Ridge Estimated Standardized Regression Coefficients for Different Biasing Constants c—Body Fat Example with Three Predictor Variables.

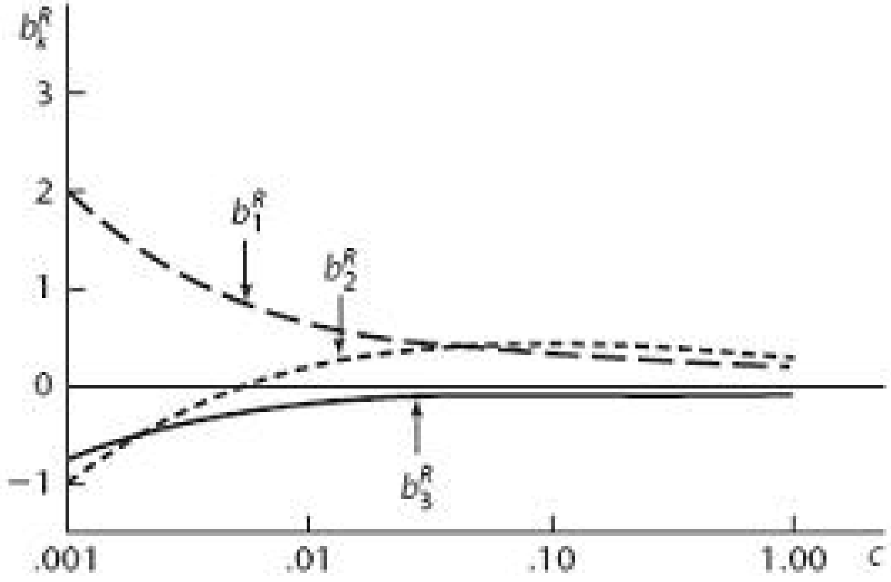
c	b_1^R	b_2^R	b_3^R
.000	4.264	−2.929	−1.561
.002	1.441	−.4113	−.4813
.004	1.006	−.0248	−.3149
.006	.8300	.1314	−.2472
.008	.7343	.2158	−.2103
.010	.6742	.2684	−.1870
.020	.5463	.3774	−.1369
.030	.5004	.4134	−.1181
.040	.4760	.4302	−.1076
.050	.4605	.4392	−.1005
.100	.4234	.4490	−.0812
.500	.3377	.3791	−.0295
1.000	.2798	.3101	−.0059

TABLE 11.3 *VIF* Values for Regression Coefficients and R^2 for Different Biasing Constants c—Body Fat Example with Three Predictor Variables.

c	$(VIF)_1$	$(VIF)_2$	$(VIF)_3$	R^2
.000	708.84	564.34	104.61	.8014
.002	50.56	40.45	8.28	.7901
.004	16.98	13.73	3.36	.7864
.006	8.50	6.98	2.19	.7847
.008	5.15	4.30	1.62	.7838
.010	3.49	2.98	1.38	.7832
.020	1.10	1.08	1.01	.7818
.030	.63	.70	.92	.7812
.040	.45	.56	.88	.7808
.050	.37	.49	.85	.7804
.100	.25	.37	.76	.7784
.500	.15	.21	.40	.7427
1.000	.11	.14	.23	.6818

Figure 5: Estimated ridge standardized regression coefficients and VIFs—body fat example

Figure 6: Ridge Trace of estimated standardized regression coefficients–bodyfat example with three predictor variables



- Note the instability of the regression coefficients for very small values of c. The estimated regression coefficient b_2^R changes signs

- It was decided to employ $c = 0.02$, because for this value of the biasing constant the ridge regression coefficients have VIF values near 1 and the estimated regression coefficients appear to have become reasonably stable.

  Choose $\lambda = 0.02$, the resulting fitted model for $c = 0.02$ is

$$\hat{Y}^* = 0.5463X_1^* + 0.3774X_2^* - 0.1369X_3^*$$

  Transforming back to the original variables by (7.53), we obtain

$$\hat{Y} = -7.3978 + 0.5553X_1 + 0.3681X_2 - 0.1917X_3$$

  where $\bar{Y} = 20.195, \bar{X}_1 = 25.305, \bar{X}_2 = 51.170, \bar{X}_3 = 27.620, s_Y = 5.106, s_1 = 5.023, s_2 = 5.235$, and $s_3 = 3.647$.

Lasso Regression (Tibshirani, 1996)

—an active area of open research

- Lasso: least absolute shrinkage and selection operator
- Automatically performs variable selection while it is estimating the regression parameters
- “Shrink” the effect of unimportant predictors, can set effects to zero
- Overall magnitude of the coefficients is constrained, important predictors are included in the model, and less important predictors shrink, potentially to zero

The least square estimates $\hat{\beta}_j$ satisfy

$\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_{p-1} x_{i,p-1})^2$

$= \text{min}_{\beta_0, \cdots, \beta_{p-1}} \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_{p-1} x_{i,p-1})^2$

Lasso: one of the various ways that one can present the lasso criterion for estimation is to minimize the least squares criterion

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_{p-1} x_{i,p-1})^2$$

subject to an upper bound on the sum of the absolute values of the regression coefficients.

$$\sum_{j=1}^{p-1} |\beta_j| \leq \lambda \sum_{j=1}^{p-1} |\hat{\beta}_j| \tag{1}$$

for some λ with $0 \leq \lambda \leq 1$

- The lasso estimates depend on the choice of λ.
- $\lambda = 1$ gives least squares estimates.
- $\lambda = 0$ gives the least squares estimates for the intercept only model

  —-i.e., it zeros out all the regression coefficients except the intercept which it estimates with $\bar{y}$

  —-all the regression coefficients in the inequality must be zero, but the intercept is not subject to the upper bound in equation 1

Cross validation:

- hold out a portion of the data (called validation set)
- fit model to the rest of the data (training set)
- determine if model based on training set performs well in validation set
- metric to assess prediction error: Mean Square Error

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2,$$

$\hat{y}_i$ is predicted value of y_i based on model.

Cross validation is used to both choose λ and assess predictive accuracy of model

- Initial training and validation sets established. Tuning parameter λ is chosen based on training set, model is fit based on training set
- Performance of the model chosen above is then assessed on the basis of the validation set
- Training model used to predict outcomes in validation set. MSE is computed. If training model produces reasonable MSE based on validation set, model is adopted.

K-fold cross validation (used to determine value of shrinkage factor λ)

- Divide data into two parts, training set and testing set
- Splits training data into $K = 10$ separate sets of equal size —-label it as $T = (T_1, T_2, \cdots, T_{10})$, training set then broke into 10 pieces
  —-commonly choose $K = 5$ or $K = 10$
- For each $k = 1, 2, \cdots, 10$, fit the model to the training set excluding the kth-fold T_k
- Compute the fitted values $\hat{y}_{i(-k)}^{(\lambda)}$ for the observations in T_k ,

based on the training data that excluded this fold

- Compute the cross-validation (CV) error for the k-th fold:

$$\big(\text{CV Error}\big)_k^{(\lambda)} = T_k^{-1} \sum_{i \in T_k} \big(y_i - \hat{y}_{i(-k)}^{(\lambda)}\big)^2$$

- The model then has overall cross-validation error:

$$\big(\text{CV Error}\big)^{\lambda} = K^{-1} \sum_{k=1}^{K} \big(\text{CV Error}\big)_k^{(\lambda)}$$

- Select λ^* as the one with minimum $\big(\text{CV Error}\big)^{\lambda}$

Example: examine the effect of lasso regression on the Coleman Report data.

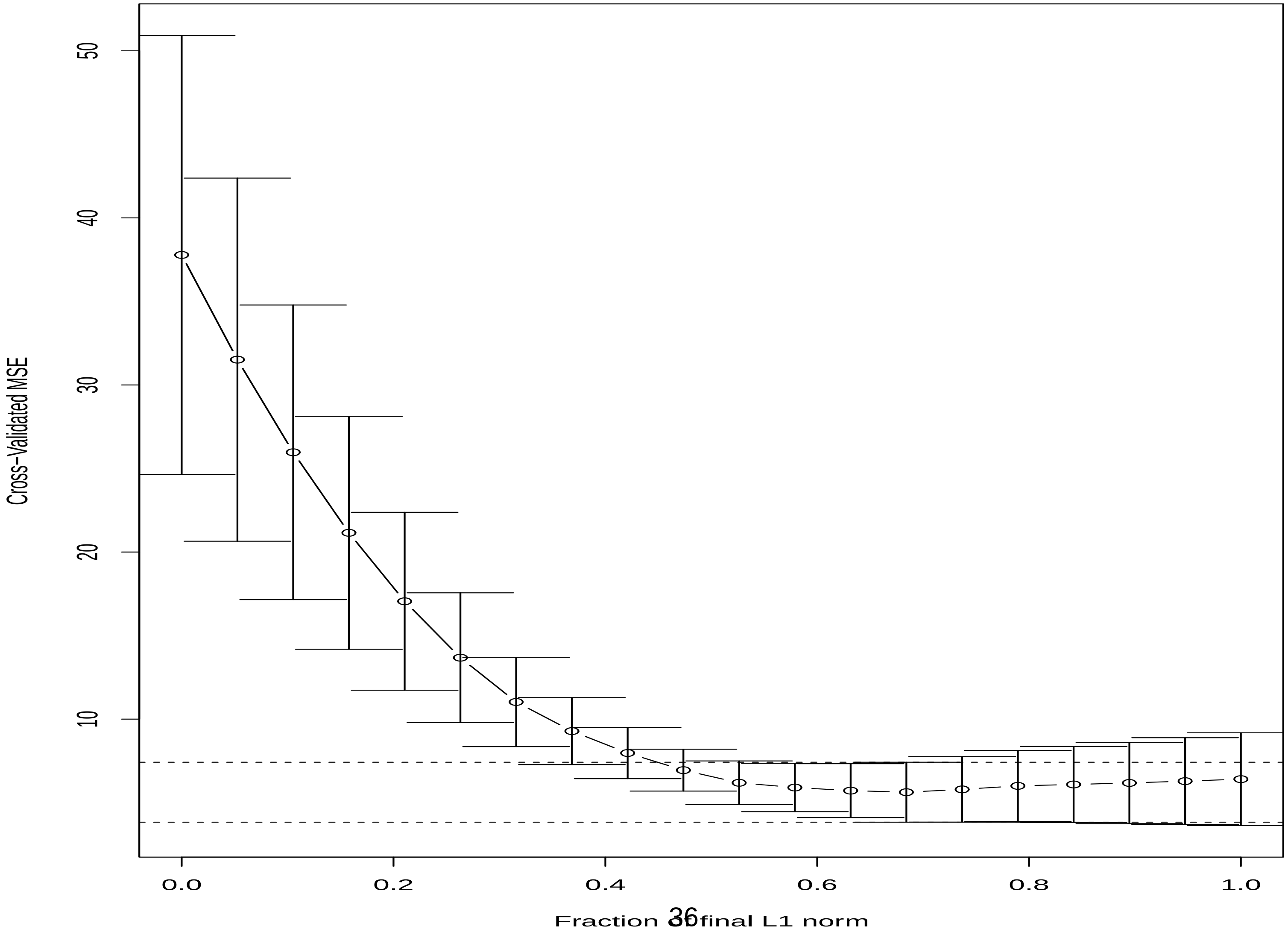
- y: the mean verbal test score for sixth graders
- x_1: staff salaries per pupil
- x_2: percentage of sixth graders whose fathers have white collar job
- x_3: a composite measure of socioeconomic status
- x_4: the mean of verbal test scores given to the teachers
- x_5: the mean educational level of the sixth grader's mothers (one unit equals two school years)

Correlation between y and the predictor variables.

	x_1	x_2	x_3	x_4	x_5
Correlation with y	0.192	0.753	0.927	0.334	0.733

- Of the five variables, x_3 has the highest correlation. It explains more of the y variable than any other single variable.
- x_2 and x_5 also have reasonably high correlations with y.
- Low correlations exist between y and both x_1 and x_4

Figure 7: CV MSE as a function of λ for coleman data, dotted line is 1 SE of lowest MSE value



	Lasso λ						Reduced	Model
Predictor	1	0.6	0.55	0.5	0.47	0	Least	Squares
Constant	19.95	18.79	21.69	26.51	23.98	35.08	12.12	14.58
x_1	-1.79	-0.34	0.00	0.00	0.00	0.00	-1.74	0.00
x_2	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00
x_3	0.56	0.52	0.50	0.48	0.49	0.00	0.55	0.54
x_4	1.11	0.62	0.47	0.28	0.38	0.00	1.04	0.75
x_5	-1.81	0.00	0.00	0.00	0.00	0.00	0.00	0.00

This Table contains results for five values of λ and least squares estimates for two reduced models

- For $\lambda = 1$, the estimates are identical to the least squares estimates for the full model

- $\lambda = 0.5$ zeros out the coefficients for x_1, x_2, and x_5.

  —-The reduced model that only includes x_3 and x_4 is the model that we liked in Chapter 9

  —-The lasso estimates of β_3 and β_4 are noticeably smaller than the least squares estimates from the reduced model given in the last column

- The largest value of λ that zeros out the coefficients x_1, x_2, and x_5 is $\lambda = 0.56348$.

  —-the lasso estimates are closer to the reduced model least squares estimates but still noticeably different

- For $\lambda \geq 0.56349$, lasso produces a nonzero coefficient for x_1. From Section 9.3, if we were going to add another variable to the model containing only x_3 and x_4, the best choice is to add x_1
- $\lambda = 0.6$ still has the coefficients for x_2 and x_5 zeroed out. The nonzero lasso estimates for β_1, β_3, and β_4 are all closer to zero than the least squares estimates from the model with just x_1, x_3, and x_4.
- Lasso seems to do a good job of identifying the important variables and it does it pretty automatically

Ridge regression and LASSO

- Ridge regression is an earlier and similar method to the lasso, and is also a shrinkage or penalization method
- Ridge regression will not set any specified predictor coefficients to exactly zero
- Lasso is preferable when predictors may be highly correlated
- For both ridge regression and lasso, λ cannot be estimated directly from the data using maximum likelihood due to an identifiability issue. This is why cross validation is chosen to fix λ at a constant