Lecture 6: Normal linear models

Maochao Xu

Department of Mathematics
Illinois State University
mxu2@ilstu.edu

Normal linear models

For GLM, assume that

$$\mathrm{E}(Y_i) = \mu_i = \mathbf{x}_i^T \boldsymbol{\beta}; \quad Y_i \sim \mathrm{N}(\mu_i, \sigma^2),$$

where $Y_1, \ldots, Y_N$ are independent random variables. The link function is the identity function, that is, $g(\mu_i) = \mu_i$. This model is usually written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

where

$$\mathbf{y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_N \end{bmatrix}, \mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_N^T \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \mathbf{e} = \begin{bmatrix} e_1 \\ \vdots \\ e_N \end{bmatrix}$$

and the e_i's are independent identically distributed random variables with $e_i \sim \mathrm{N}(0, \sigma^2)$ for $i = 1, \ldots, N$.

Multiple linear regression, analysis of vari-ance (ANOVA) and analysis of covariance (ANCOVA) are all of this form and together are sometimes called **general linear models**.

Specific form

The general multiple linear regression model with response Y and terms $X_1, \ldots, X_p$ will have the form

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_{p-1} X_{p-1} + e.$$

Suppose we have observed data for N cases or units, meaning we have a value of Y and all of the terms for each of the N cases.

cases	Y	X_1	X_2	$\ldots$	X_{p-1}
1	Y_1	X_{11}	X_{12}	$\ldots$	$X_{1,p-1}$
2	Y_2	X_{21}	X_{22}	$\ldots$	$X_{2,p-1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	Y_N	X_{N1}	X_{N2}	$\ldots$	$X_{N,p-1}$

Note that

$$\mathbf{X} = \begin{pmatrix} 1 & X_{11} & X_{12} & \ldots & X_{1,p-1} \\ \vdots & & & & \\ 1 & X_{N1} & X_{N2} & \ldots & X_{N,p-1} \end{pmatrix}.$$

Estimation: MLE

the maximum likelihood estimator of $\boldsymbol{\beta}$ is given by

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y},$$

provided $(\mathbf{X}^T\mathbf{X})$ is non-singular. As $\mathrm{E}(\mathbf{b}) = \boldsymbol{\beta}$, the estimator is unbiased. It has variance–covariance matrix $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1} = \mathfrak{I}^{-1}$.

In the context of generalized linear models, σ^2 is treated as a nuisance parameter. However, it can be shown that

$$\widehat{\sigma}^2 = \frac{1}{N-p}(\mathbf{y}-\mathbf{X}\mathbf{b})^T(\mathbf{y}-\mathbf{X}\mathbf{b}) = \text{MSE}$$

is an unbiased estimator of σ^2

(MSE=mean square error)

this can be used to estimate $\mathfrak{I}$ and hence make inferences about $\mathbf{b}$.

Estimation: Least squares estimation

The least squares estimate is chosen to minimize the residual sum of squares function

$$RSS(\beta) = \min_{\beta_1,\ldots,\beta_p} \sum_{i=1}^{N} (Y_i - \mathbf{X_i}^T\beta)^2 = (\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta).$$

Taking the derivative,

$$\frac{\partial Q}{\partial \beta} = -2\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\beta),$$

we have the following normal equation

$$\mathbf{X}^T\mathbf{Y} = \mathbf{X}^T\mathbf{X}\beta.$$

The estimate for β is

$$\boldsymbol{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

if the inverse exists.

- Fitted values:
$$\hat{\mathbf{Y}} = \mathbf{Xb}.$$
- Residuals:
$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}.$$
Note that
$$\hat{\mathbf{Y}} = \mathbf{HY},$$
where
$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T,$$
is the hat matrix. Similarly,
$$\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{Y}.$$
The variance-covariance matrix of the residuals is
$$\mathrm{Var}(\mathbf{e}) = \sigma^2(\mathbf{I} - \mathbf{H}).$$
- The residual vector $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$ is orthogonal to the subspace. $\hat{\mathbf{Y}}$ is the projection of $\mathbf{Y}$ on the subspace.
- The hat matrix $\mathbf{H}$ computes the orthogonal projection, and hence it is also known as a projection matrix.

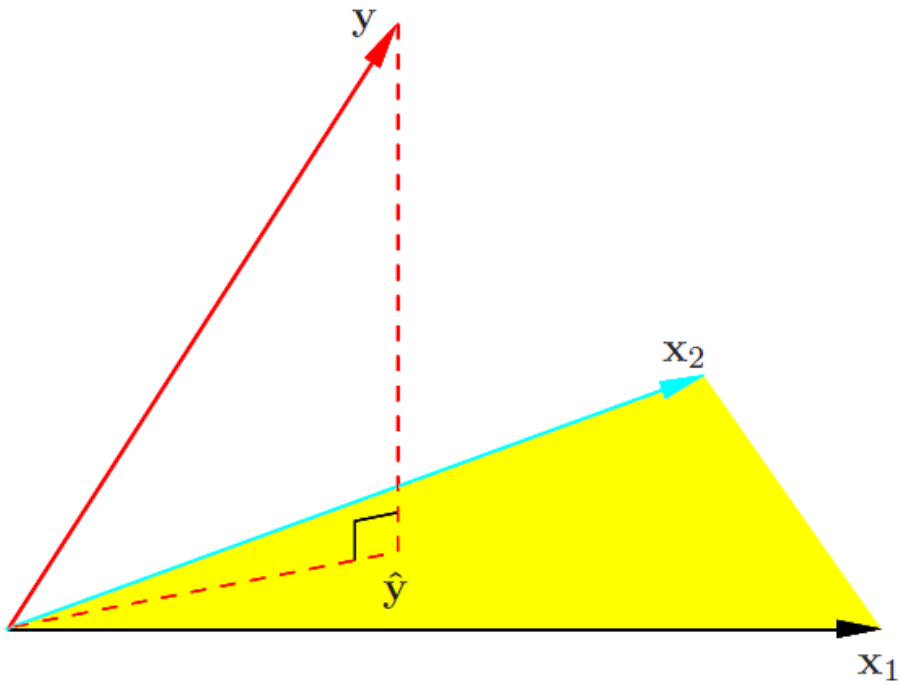


FIGURE 3.2. *The N-dimensional geometry of least squares regression with two predictors. The outcome vector* $\mathbf{y}$ *is orthogonally projected onto the hyperplane spanned by the input vectors* $\mathbf{x}_1$ *and* $\mathbf{x}_2$*. The projection* $\hat{\mathbf{y}}$ *represents the vector of the least squares predictions*

Hypothesis testing

Consider a null hypothesis H_0 and a more general hypothesis H_1 specified as

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0 = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_q \end{bmatrix} \quad \text{and} \quad H_1 : \boldsymbol{\beta} = \boldsymbol{\beta}_1 = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix},$$

where $q < p < N$. Let $\mathbf{X}_0$ and $\mathbf{X}_1$ denote the corresponding design matrices, $\mathbf{b}_0$ and $\mathbf{b}_1$ the maximum likelihood estimators, and D_0 and D_1 the deviances.

Therefore, the statistic

$$F = \frac{D_0 - D_1}{p-q} \Big/ \frac{D_1}{N-p} = \frac{(\mathbf{b}_1^T \mathbf{X}_1^T \mathbf{y} - \mathbf{b}_0^T \mathbf{X}_0^T \mathbf{y})}{p-q} \Big/ \frac{(\mathbf{y}^T \mathbf{y} - \mathbf{b}_1^T \mathbf{X}_1^T \mathbf{y})}{N-p}$$

will have the central distribution $F(p-q, N-p)$ if H_0 is correct; F will otherwise have a non-central distribution. Therefore, values of F that are large relative to the distribution $F(p-q, N-p)$ provide evidence against H_0.

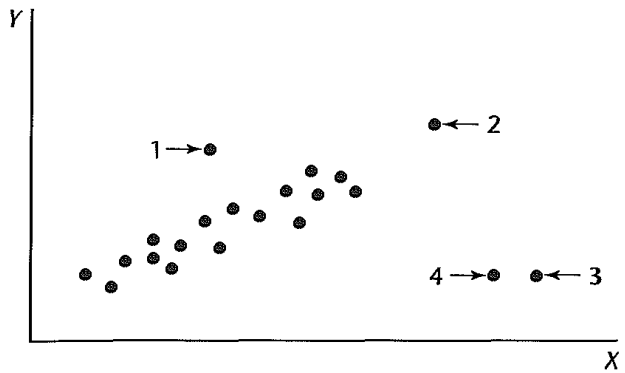
An alternative view of the F statistic is in terms of the residual sums of squares

$$F = \frac{S_0 - S_1}{p-q} \Big/ \frac{S_1}{N-p},$$

where $S_0 = \mathbf{y}^T \mathbf{y} - \mathbf{b}_0^T \mathbf{X}_0^T \mathbf{y}$ and $S_1 = \mathbf{y}^T \mathbf{y} - \mathbf{b}_1^T \mathbf{X}_1^T \mathbf{y}$.

Outlying cases

It is important to study the outlying cases and decide whether they should be retained or eliminated, and if retained, whether their influence should be reduced in the fitting process and/or the regression model should be revised. **Not all outlying cases have a strong influence on the fitted regression function.**



Which points are influence?

Residuals

$$e_i = Y_i - \hat{Y}_i.$$

Semistudentized residuals

$$e_i^* = \frac{e_i}{\sqrt{MSE}}.$$

Studentized Residuals and deleted Residuals

Note that

$$s(e_i) = \sqrt{MSE(1 - h_{ii})}.$$

- The Studentized Residual (standardized residuals) is defined as

$$r_i = \frac{e_i}{s(e_i)}.$$

- Deleted Residuals: measure the ith residual $e_i = Y_i - \hat{Y}_i$ when the fitted regression is based on all the cases except the ith one.

  Procedure: delete the ith case, fit the regression function to the remaining $n - 1$ cases, and obtain the point estimate of the expected value when the X levels are those of the ith case, to be denoted by $\hat{Y}_{i(i)}$. The difference between the actual observed value Y_i and the estimated expected value $\hat{Y}_{i(i)}$ will be denote by d_i

$$d_i = Y_i - \hat{Y}_{i(i)}$$

  the difference d_i is called the **deleted residual** for the ith case. In fact, we can prove that

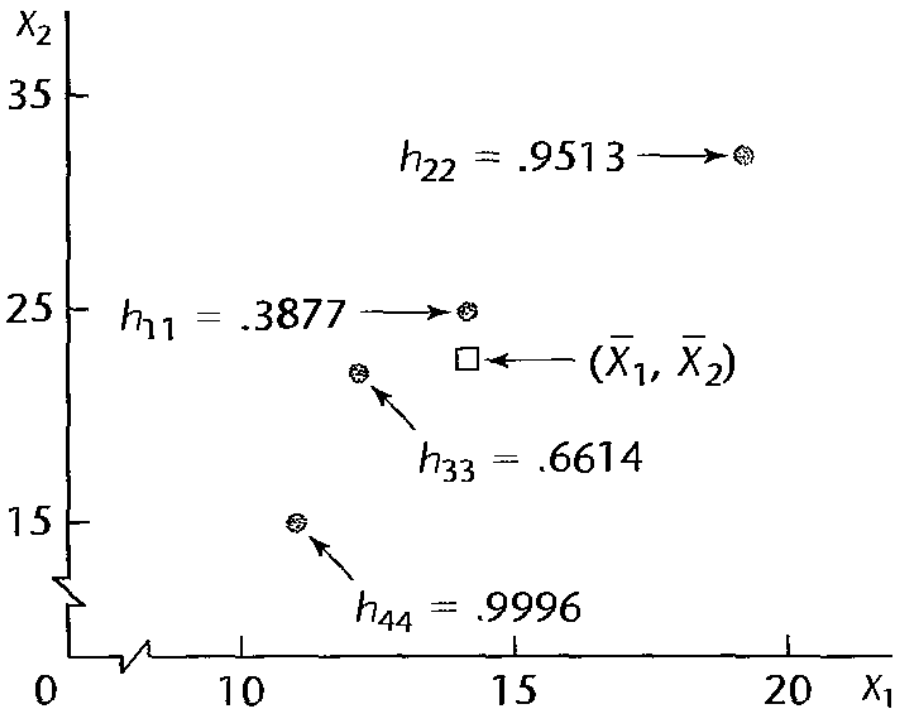
$$d_i = \frac{e_i}{1 - h_{ii}}.$$

Hat matrix: leverage values

The diagonal elements h_{ii} have the following properties

$$0 \leq h_{ii} \leq 1, \quad \sum_{i=1}^{n} h_{ii} = p.$$

It can be shown that h_{ii} is a measure of the distance between the X values for the ith case and the means of the X values for all n cases. Thus, a large value h_{ii} indicates the ith case is distant from the center of all X observations, which is called **leverage** of the ith case.



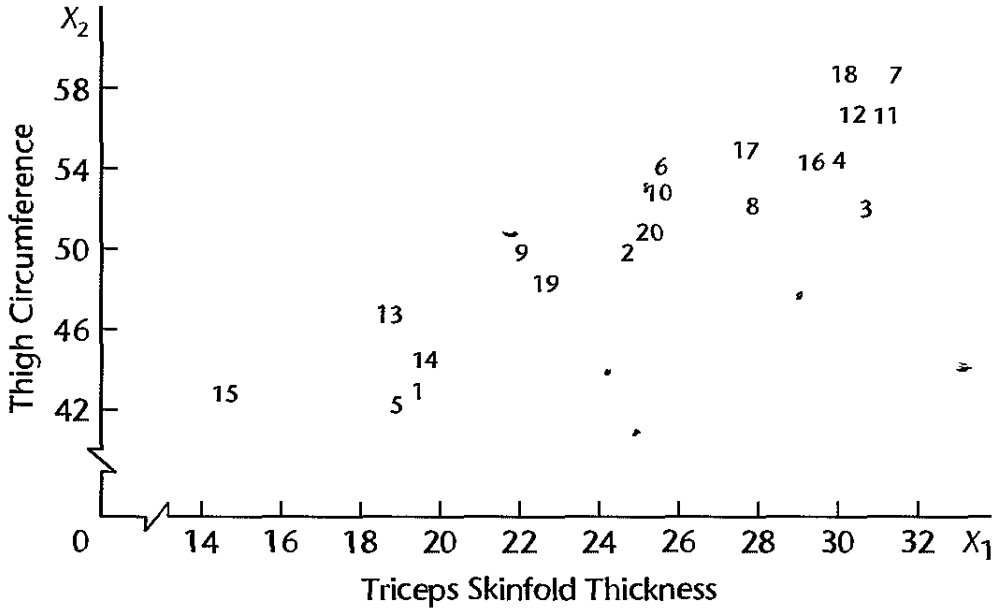
If the ith case is outlying in terms of its X observations and therefore has a large leverage value h_{ii}, it exercises substantial leverage in determining the fitted value $\hat{Y}_i$.

- h_{ii} is the weight of observation Y_i in determining the fitted value.
- The larger is h_{ii}, the smaller is the variance of the residual e_i. Hence, the larger is h_{ii}, the closer the fitted value $\hat{Y}_i$ will tend to be the observed value Y_i.

A leverage value is considered to be large if it is more than twice as large as the mean leverage value

$$\bar{h} = \frac{\sum_{i=1}^{n} h_{ii}}{n} = \frac{p}{n}.$$

Hence, leverage values greater than $2p/n$ are considered by this rule to indicate outlying cases with regard to their X values. The other guideline is that h_{ii} values exceeding .5 indicates very high leverage, whereas those between .2 and .5 indicate moderate leverage.



Influential cases

We consider a case to be *influential* if its exclusion causes major changes in the fitted regression function.

DFFITS: A measure of the influence that case i has on the fitted value $\hat{Y}_i$ is given by

$$(DFFITS)_i = \frac{\hat{Y} - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)} h_{ii}}}.$$

It can be shown that

$$(DFFITS)_i = r_i \left(\frac{h_{ii}}{1 - h_{ii}} \right)^{1/2}.$$

Guideline:

$$(DFFITS)_i > 1$$

for small to medium data sets, and

$$(DFFITS)_i > 2\sqrt{p/n}$$

for large data sets.

Cook's Distance

In contrast to DFFITS, Cook's distance measure is an aggregate influence measure, showing the effect of the i case on all n fitted values:

$$D_i = \frac{\sum_{j=1}^{n} \left(\hat{Y}_j - \hat{Y}_{j(i)} \right)^2}{p \cdot MSE}.$$

We can prove that

$$D_i = \frac{r_i^2}{p} \left[\frac{h_{ii}}{1 - h_{ii})} \right].$$

Guideline: If the percentile value is less than 10 or 20 percent, the ith case has little apparent influence on the fitted values. On the other hand, if the percentile value is near 50 percent or more, the fitted values obtained with and without the ith case should be considered to differ substantially.

DFBETAS: Influence on the regression coefficients

A measure of the influence of the ith case on each regression coefficient b_k is the difference between the estimated regression coefficient b_k based on all n cases and the regression coefficient obtained when the ith case is omitted, denoted by $b_{k(i)}$.

$$(DFBETAS)_{k(i)} = \frac{b_k - b_{k(i)}}{\sqrt{MSE_{(i)}\, c_{kk}}}$$

where c_{kk} is the kth diagonal element of $(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}$.

Guideline: The absolute value of DFBETAS exceeds 1 for small medium data sets and $2/\sqrt{n}$ for large data sets.

	(1)	(2)	(3)	(4)	(5)
				DFBETAS	
i	$(DFFITS)_i$	D_i	b_0	b_1	b_2
1	−.366	.046	−.305	−.132	.232
2	.384	.046	.173	.115	−.143
3	−1.273	.490	−.847	−1.183	1.067
4	−.476	.072	−.102	−.294	.196
5	.000	.000	.000	.000	.000
6	−.057	.001	.040	.040	−.044
7	.128	.006	−.078	−.016	.054
8	.575	.098	.261	.391	−.333
9	.402	.053	−.151	−.295	.247
10	−.364	.044	.238	.245	−.269
11	.051	.001	−.009	.017	−.003
12	.323	.035	−.131	.023	.070
13	−.851	.212	.119	.592	−.390
14	.636	.125	.452	.113	−.298
15	.189	.013	−.003	−.125	.069
16	.084	.002	.009	.043	−.025
17	−.118	.005	.080	.055	−.076
18	−.166	.010	.132	.075	−.116
19	−.315	.032	−.130	−.004	.064
20	.094	.003	.010	.002	−.003

Coefficient of Multiple Determination

The coefficient of multiple determination is defined as

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO},$$

which measures the proportionate reduction of total variation in Y associated with the use of the set of $\mathbf{X}$ variables.

Note that adding more X variables to the regression model can only *increase* R^2 and never reduce it (why?). Hence, the following *adjusted coefficient of multiple determination*, denoted by R^2_a, is defined as

$$R^2_a = 1 - \frac{SSE/(n-p)}{SSTO/(n-1)} = 1 - \frac{(n-1)}{(n-p)}\frac{SSE}{SSTO}.$$

Multicollinearity: Variance Inflation Factor

Indications of the presence of serious multicollinearity are given by the informal diagnostics:

- Large changes in the estimated regression coefficients when a predictor variable is added or deleted, or when an observation is altered or deleted.
- Nonsignificant results in individual tests on the regression coefficients for important predictor variables.
- Estimated regression coefficients with an algebraic sign that is the opposite of that expected from theoretical considerations or prior experience.
- Large coefficients of simple correlation between pairs of predictor variables in the correlation matrix.
- Wide confidence intervals for the regression coefficients representing important predictor variables.

Variance inflation factor

Collinearity can be detected by calculating the variance inflation factor (VIF). The $(VIF)_k$ is called the **variance inflation factor** for b_k. It can be shown that

$$(VIF)_k = \left(1 - R_k^2\right)^{-1}, \quad k = 1, 2, \ldots, p-1,$$

where R_k^2 is the coefficient of multiple determination when X_k is regressed on the other variables in the model.

The largest VIF value is often used as an indicator of the severity of multicollinearity. A maximum VIF value in excess of 10 is frequently taken as an indication that multicollinearity may be unduly influencing the least squares estimates.

The mean VIF is

$$(\overline{VIF}) = \frac{\sum_{k=1}^{p-1}(VIF)_k}{p-1}.$$

Mean VIF values considerably larger than 1 are indicative of serious multicollinearity.

Model selection

Akaike's information criterion (AIC) can be motivated in two ways. The most popular motivation seems to be based on balancing goodness of fit and a penalty for model complexity. AIC is defined such that *the smaller the value of AIC the better the model.* A measure of goodness of fit such that the smaller the better is minus one times the likelihood associated with the fitted model, while a measure of complexity is m, the number of estimated parameters in the fitted model.

$$AIC = n\log(SSE_m/n) + 2m.$$

Recall that m is the number of parameters in your subset model. For example, if your model includes only $\beta_0, \beta_1, \beta_3$, then $m = 3$.

Caution: When the sample size is small, or when the number of parameters estimated is a moderate to large fraction of the sample size, it is well-known that AIC has a tendency for over-fitting since the penalty for model complexity is not strong enough.

BIC: Bayes Information Criterion

BIC is defined such that the smaller the value of BIC the better the model.

$$BIC = n \log(SSE_m/n) + m log(n).$$

BIC penalizes complex models more heavily than AIC, thus favoring simpler models than AIC.

Simonoff (2003, p. 46) concludes the following:

AIC has the desirable property that it is an efficient model selection criterion. What this means is that as the sample gets larger, the error obtained in making predictions using the model chosen using the AIC becomes indistinguishable from the error obtained using the best possible model among all candidate models. That is, in this large-sample predictive sense, it is as if the best approximation was known to the data analyst. Other criteria, such as the BIC do not have this property.

Hastie, Tibshirani and Freedman (2001, p. 208) put forward the following different point of view:

For model selection purposes, there is no clear choice between AIC and BIC. BIC is asymptotically consistent as a selection criterion. What this means is that given a family of models, including the true model, the probability that BIC will select the correct model approaches one as the sample size $N \to \infty$. This is not the case for AIC, which tends to choose models which are too complex as $N \to \infty$. On the other hand, for finite samples, BIC often chooses models that are too simple, because of the heavy penalty on complexity.

Stepwise procedures

Suppose our model:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + e.$$

Forward selection uses the following procedure:

- Fit a simple linear regression for each X_i. The X_i variable with largest t value is the candidate for first addition. If this t value exceeds a predetermined level α, the X is added. Otherwise, stop.
- Assume X_2 is the variable entered at step 1. Now, fit all regression models with two (X_j, X_2) variables, where X_2 is one of the pair. Check the t value for the new variable X_j. If it is large, keep it, otherwise, drop it, and stop.
- Suppose X_3 is added at the above stage. Now, check whether any of the other X variables already in the model should be dropped. That is, check the t value for X_2 in the new model. If it is large, keep it, otherwise, drop it.
- Suppose X_2 is retained, so that X_2 and X_3 are both in the model. Continue to examine to which X variable is the next candidate for addition, then examine whether any of the variables already in the model should be dropped.
- The routine stops when there is no further X variables can be added or deleted.

Backward elimination algorithm

The *backward elimination algorithm* works in the opposite order:

- Fit the model with all X variables. If the largest p-value for X_i exceeds a predetermined limit, then the variable is dropped.
- Refitting the model with the rest of X variables, and the next candidate for dropping is identified.
- The process continues until no further X variables can be dropped.

The forward and backward algorithms can be combined into a stepwise method, where at each step, a term is either deleted or added so that the resulting candidate mean function minimizes the criterion function of interest.

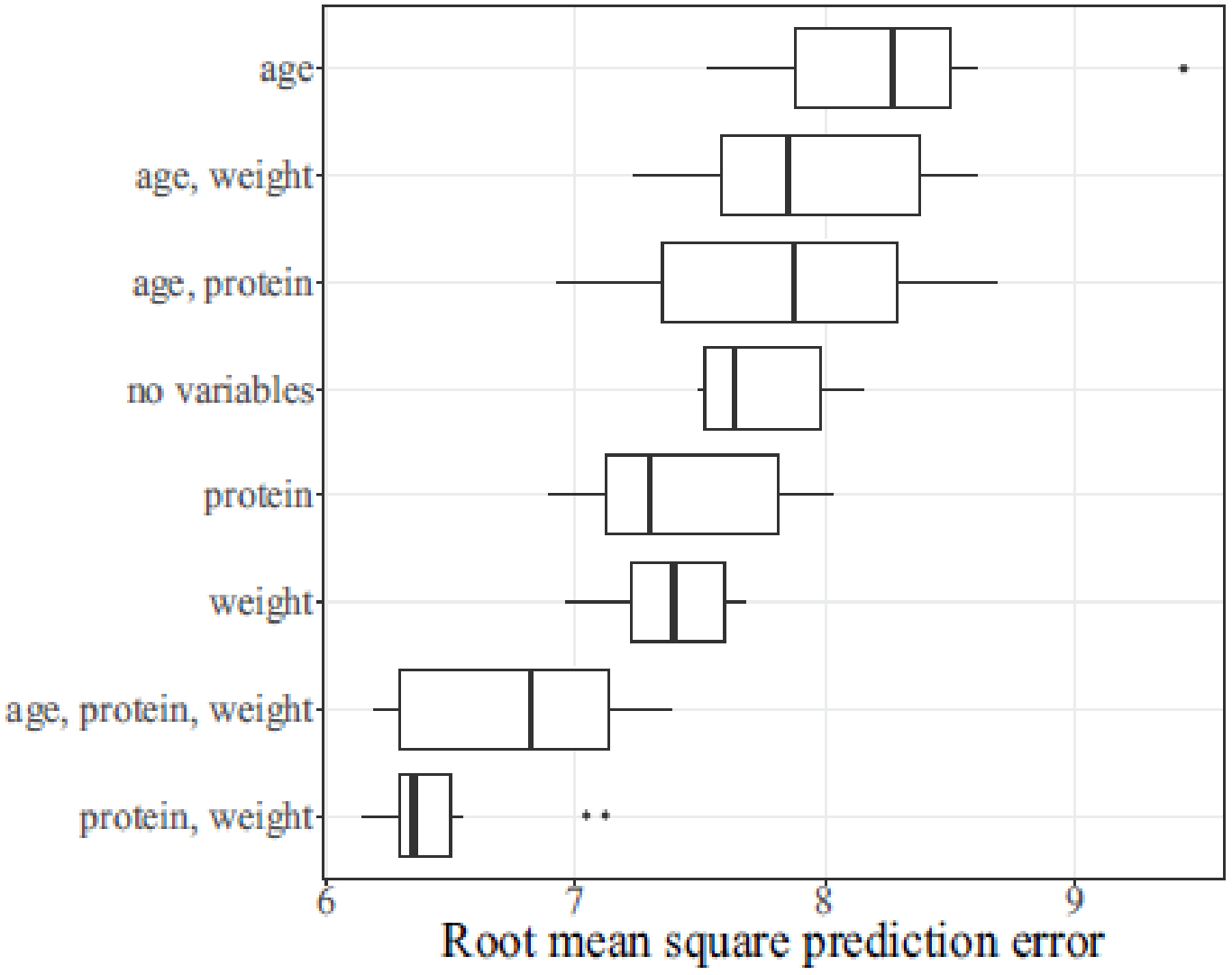
Cross-validation

For k-fold cross-validation the sample is randomly split into k non-overlapping samples or "folds" of size $m = n/k$. The k-fold cross-validation can also be replicated multiple times to further increase robustness.

Cross-validation can be used to choose between two alternative models, or where the number of explanatory variables is reasonably small it is fea-sible to use an exhaustive search over all possible models.

For three explana-tory variables there are eight different combinations of variables ranging from none included to all three included. The root mean square prediction errors were based on 5-fold cross-validation replicated 10 times. The model with the lowest errors has the two explanatory variables, protein and weight, and this model generally has lower errors than the full model which also includes age. The three models with the worst prediction error all include age, suggesting that age is not a useful explanatory variable.

Cross-validation



Ridge Regression

Recall that the least squares fitting procedure minimizes the residual sum of squares

$$\begin{aligned} RSS(\beta) &= \sum_{i=1}^{N}(y_i - f(\mathbf{x_i}))^2 \\ &= \sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2. \end{aligned}$$

Ridge regression minimizes a slightly different quantity:

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

where $\lambda \geq 0$ is a tuning parameter, to be determined separately. The term $\lambda\sum_{j=1}^{p}\beta_j^2$ is called a **shrinkage penalty**, which has the effect of shrinking the estimates β_j towards zero. Unlike least squares, which generates only one set of coefficient estimates, ridge regression will produce a different set of coefficient estimates for each value of λ. Selecting a good value for λ is critical.

An equivalent way to write the ridge problem is

$$\hat{\beta}^{ridge} = \arg\min \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2$$

subject to

$$\sum_{j=1}^{p} \beta_j^2 \leq t,$$

which makes explicit the size constraint on the parameters. There is a one-to-one correspondence between the parameters λ and t.

Note that the shrinkage penalty is applied to β_j, $1 \leq j \leq p$, but not to the intercept β_0. We want to shrink the estimated association of each variable with the response; however, we do not want to shrink the intercept, which is simply a measure of the mean value of the response when $x_{ij} = 0, j = 1, \ldots, p$. If we assume that the variables-that is, the columns of the data matrix $\mathbf{X}$-have been centered to have mean zero before ridge regression is performed, then the estimated intercept will take the form

$$\hat{\beta}_0 = \sum_{i=1}^{N} y_i / N = \bar{y}.$$

Ridge Regression

First, we need to center the variables of $\mathbf{X}$, that is, replacing each x_{ij} by

$$x_{ij} - \bar{x}_j.$$

Then, β_0 is estimated by

$$\hat{\beta}_0 = \sum_{i=1}^{N} y_i / N.$$

The remaining coefficients get estimated by a ridge regression without intercept, using the centered x_{ij}. Henceforth we assume that this centering has been done, so that the input matrix $\mathbf{X}$ has p (rather than $p+1$) columns.

Writing the criterion of ridge regression in matrix form

$$RSS(\lambda) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^T\boldsymbol{\beta},$$

we have he following solution

$$\hat{\boldsymbol{\beta}}^{ridge} = \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{y}.$$

The solution adds a positive constant to the diagonal of $\mathbf{X}^T\mathbf{X}$ before inversion. This makes the problem nonsingular, even if $\mathbf{X}^T\mathbf{X}$ **is not of full rank**.

Comparison: Ridge and OLS

Why Does Ridge Regression Improve Over Least Squares? Ridge regression's advantage over least squares is rooted in the **bias-variance** trade-off. As λ increases, the flexibility of the ridge regression fit decreases, leading to decreased variance but increased bias.

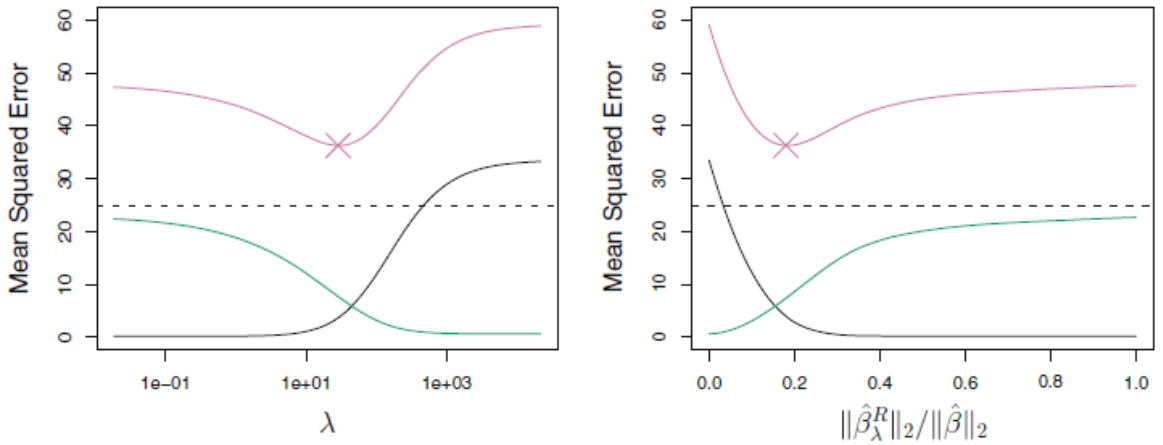


FIGURE *Squared bias (black), variance (green), and test mean squared error (purple) for the ridge regression predictions on a simulated data set, as a function of λ and $\|\hat{\beta}^R_\lambda\|_2/\|\hat{\beta}\|_2$. The horizontal dashed lines indicate the minimum possible MSE. The purple crosses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

In general, in situations where the relationship between the response and the predictors is close to linear, the least squares estimates will have low bias but may have high variance. This means that a small change in the training data can cause a large change in the least squares coefficient estimates. **Ridge regression works best in situations where the least squares estimates have high variance.**

Ridge regression also has substantial computational advantages over best subset selection, which requires searching through 2^p models.

For the special case of an orthonormal design matrix, i.e.,

$$\mathbf{X}\mathbf{X}^T = \mathbf{I},$$

we have

$$\hat{\boldsymbol{\beta}}^{ridge} = \frac{1}{1+\lambda}\boldsymbol{\beta}^{ols}.$$

What you could observe from here?

Some theoretical results:

- The total variance

$$\sum_{j=1}^{p} \mathrm{Var}\left(\hat{\beta}_j^{ridge}\right)$$

  is a monotone decreasing sequence with respect to λ.

- The total bias

$$\sum_{j=1}^{p} \mathrm{Bias}^2\left(\hat{\beta}_j^{ridge}\right)$$

  is a monotone increasing sequence with respect to λ.

LASSO

The lasso is a shrinkage method like ridge, which minimizes

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

In statistical, the lasso uses an l_1 penalty instead of an l_2 penalty. Equivalently, it can be rewritten as

$$\hat{\beta}^{lasso}=\arg\min\sum_{i=1}^{N}\left(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2$$

subject to

$$\sum_{j=1}^{p}|\beta_j|\leq t.$$

In the signal processing literature, the lasso is also known as **basis pursuit**.

As with ridge regression, the lasso shrinks the coefficient estimates towards zero. However, in the case of the lasso, **the l_1 penalty has the effect of forcing some of the coefficient estimates to be exactly equal to zero** when the tuning parameter λ is sufficiently large. Hence, much like best subset selection, the lasso performs variable selection. As a result, models generated from the lasso are generally much easier to interpret than those produced by ridge regression. We say that the lasso yields sparse models-that is, sparse models that involve only a subset of the variables. As in ridge regression, selecting a good value of λ for the lasso is critical.

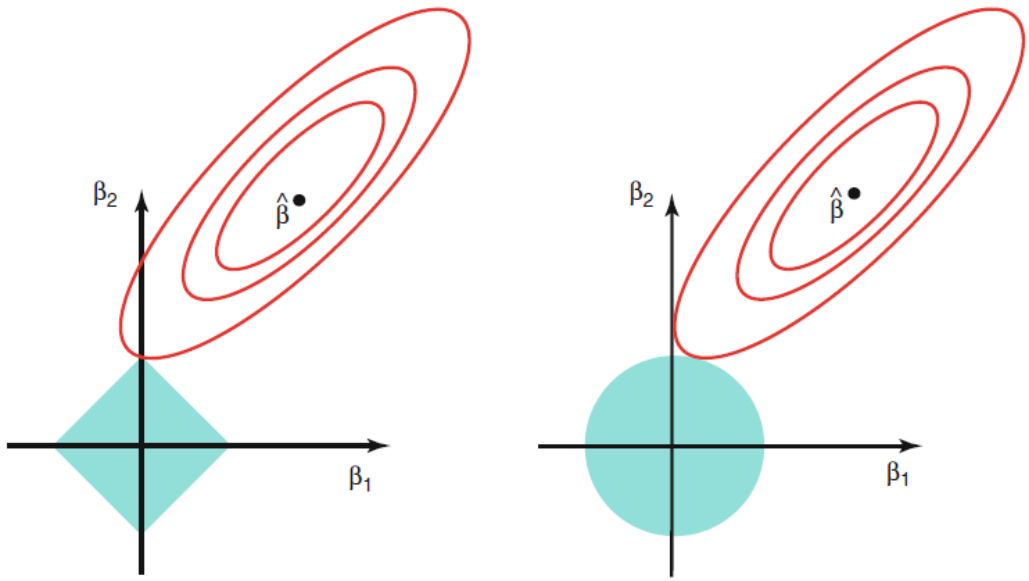


FIGURE *Contours of the error and constraint functions for the lasso* (left) *and ridge regression* (right). *The solid blue areas are the constraint regions,* $|\beta_1| + |\beta_2| \leq s$ *and* $\beta_1^2 + \beta_2^2 \leq s$, *while the red ellipses are the contours of the RSS.*

Simple Example

Assume that for the quantity

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

we have

$$\beta_0=0,\quad x_{ii}=1,\quad x_{ij}=0, i\neq j.$$

Then, for OLS, we have

$$\sum_{i=1}^{p}(y_i-\beta_i)^2\,.$$

In this case, the least squares solution is given by

$$\hat{\beta}_i=y_i.$$

For the ridge regression, we minimize

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

The lasso amounts to finding the coefficients such that

$$\sum_{i=1}^{p}(y_i-\beta_i)^2+\lambda\sum_{j=1}^{p}|\beta_j|$$

is minimized.

The ridge regression estimates take the form

$$\hat{\beta}_j^{ridge} = y_j/(1+\lambda).$$

The lasso estimates take the form

$$\hat{\beta}_j^{lasso} = \begin{cases} y_j - .5\lambda, & y_j > .5\lambda \\ y_j + .5\lambda, & y_j < -.5\lambda \\ 0, & |y_j| \leq .5\lambda. \end{cases}$$

In ridge regression, each least squares coefficient estimate is shrunken by the same proportion. In contrast, the lasso shrinks each least squares coefficient towards zero by a constant amount $\lambda/2$.

L_q Penalty

The penalty function could be very flexible.

$$\lambda \|\boldsymbol{\beta}\|_q^q = \lambda \sum_{j=1}^{d} |\beta_j|^q, q \geq 0.$$

- LASSO: $q = 1$

$$\sum_{j=1}^{d} |\beta_j|$$

- Ridge: $q = 2$

$$\sum_{j=1}^{d} \beta_j^2$$

- Supnorm: $q = \infty$

$$\max_j |\beta_j|.$$

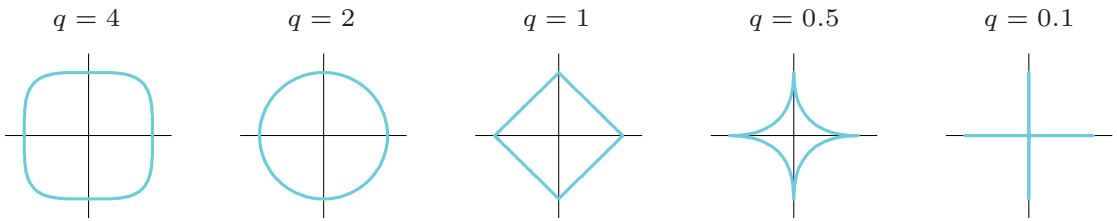


FIGURE 3.12. *Contours of constant value of* $\sum_j |\beta_j|^q$ *for given values of* q.

- $q \leq 1$ more weights imposed on the coordinate directions.
- $q = 1$ is the smallest q such that the constraint region is convex
- q could be fractions.

Soft Thresholding

For the orthonormal case,

$$\beta_j^{lasso} = \text{sign}\left(\hat{\beta}_j\right)\left(\|\hat{\beta}_j\| - \lambda/2\right)_+ .$$

That is, if $\hat{\beta}_j > \lambda/2$,

$$\beta_j^{lasso} = \hat{\beta}_j - \lambda/2;$$

if $\hat{\beta}_j \leq \lambda/2$,

$$\beta_j^{lasso} = 0;$$

if $\hat{\beta}_j < -\lambda/2$,

$$\beta_j^{lasso} = \hat{\beta}_j + \lambda/2.$$

- The Lasso shrinks big coefficients by a constant,
- Lasso truncates small coefficients to zero.

Elastic net: (Zou and Hastie 2005)

$$\lambda \sum_{j=1}^{p} \left[\alpha \beta_j^2 + (1-\alpha)|\beta_j| \right],$$

a compromise between ridge and lasso. The elastic-net selects variables like the lasso, and shrinks together the coefficients of correlated predictors like ridge.

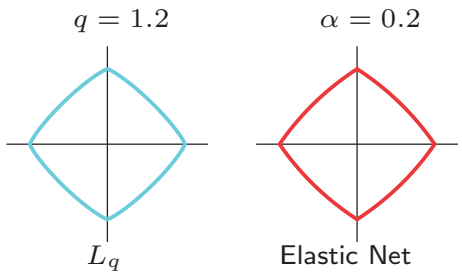


FIGURE 3.13. *Contours of constant value of* $\sum_j |\beta_j|^q$ *for* $q = 1.2$ *(left plot), and the elastic-net penalty* $\sum_j (\alpha\beta_j^2 + (1-\alpha)|\beta_j|)$ *for* $\alpha = 0.2$ *(right plot). Although visually very similar, the elastic-net has sharp (non-differentiable) corners, while the* $q = 1.2$ *penalty does not.*