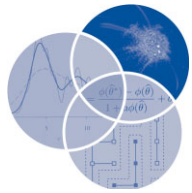
Overview



Linear regression

Xiaogang Su,[1]* Xin Yan[2] and Chih-Ling Tsai[3]

Linear regression plays a fundamental role in statistical modeling. This article provides a step-by-step coverage of linear models in the order of model specification, model estimation, statistical inference, variable selection, model diagnosis, and prediction. Computation issues in linear regression and intimately relevant extensions of linear models are also discussed. © 2012 Wiley Periodicals, Inc.

How to cite this article:
WIREs Comput Stat 2012, 4:275–294. doi: 10.1002/wics.1198

Keywords: linear regression; model diagnosis; shrinkage; statistical inference; variable selection

INTRODUCTION

Consider the regression problem in which a continuous response Y is to be regressed on a number of predictors $X_1, \ldots, X_p$. It is known that linear regression provides the simplest model form to model the regression function as a linear combination of predictors. It is popular in applications, and several reasons account for its popularity given below. Because of the linear form, the model parameters are easily interpretable. In addition, linear model theories are well established with mathematical elegance. Moreover, linear regression is the building block for many modern modeling tools. In particular, when the sample size is small or the signal is relatively weak, linear regression often provides a satisfactory approximation to the underlying regression function.

This article provides a concise account of major aspects involved in linear regression. The exposition follows the natural flow in typical model fitting, which includes model specification, least squares estimation, statistical inference, model selection, model diagnostics, and model deployment or prediction. We also discuss computational issues and some relevant extensions. In the *Conclusion* section, we briefly summarize and discuss the extensions that have been omitted from this coverage.

*Correspondence to: xgsu@uab.edu

[1]School of Nursing, University of Alabama, Birmingham, AL, USA

[2]Department of Statistics, University of Central Florida, Orlando, FL, USA

[3]Graduate School of Management, University of California, Davis, CA, USA

MODEL SPECIFICATION

Consider data $\mathcal{D} = \{(y_i, \mathbf{x}_i) : i = 1, \ldots, n\}$, where y_i is the ith response, measured on a continuous scale; $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{t}} \in \mathbb{R}^p$ is the associated predictor vector; and n ($\gg p$) is the sample size. The linear model is specified as

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i \quad \text{with } \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma^2\right), \tag{1}$$

for $i = 1, \ldots, n$. In matrix form,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad \text{with } \boldsymbol{\varepsilon} \sim \mathcal{N}\left(0, \sigma^2 \mathbf{I}\right), \tag{2}$$

where $\mathbf{y} = [y_i]_{n\times 1}$ is the n-dimensional response vector; $\mathbf{X} = \left(x_{ij}\right)_{n\times(p+1)}$ with $x_{i0} = 1$ is often called the design matrix; and $\boldsymbol{\varepsilon} = [\varepsilon_i]_{n\times 1}$. There are four major statistical assumptions involved in the specification of model (1) or (2), and they are

1. (Linearity) $\boldsymbol{\mu} \equiv [E(y_i|\mathbf{x}_i)]_{n\times 1} = \mathbf{X}\boldsymbol{\beta}$;
2. (Independence) ε_i's are independent of each other;
3. (Homoscedasticity) ε_i's have equal variance σ^2;
4. (Normality) ε_i's are normally distributed.

It is noteworthy that many properties of linear models remain valid without all four assumptions. However, we make no effort in elaborating these details in this article. To extract model interpretation,

 © 2012 Wiley Periodicals, Inc.

we introduce a generic notation μ_x to denote the conditional mean response as

$$\mu_x = E(Y|X_1, \ldots X_p) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p,$$

where X_j is an $n \times 1$ vector for $j = 1, \ldots, p$. As $\partial \mu_x / \partial X_j = \beta_j$, the regression parameters can be easily interpreted in terms of change rate. That is, β_j corresponds to the amount of change in the conditional mean response μ_x with one unit increase in X_j, given all other predictors are fixed. To illustrate, consider a practical example by regressing systolic blood pressure (SBP) on age, race, and gender. The slope parameter β for age may be interpreted using the following statement. Given two individuals A and B where both are of the same race and gender, but A is $\tilde{a}$ years older than B, the SBP level of A is expected to be $(\tilde{a}\beta)$ higher than that of B.

The above model specification is flexible enough to incorporate the following three important scenarios. First, interaction terms can be included as cross products, for example, $X_1 X_2$ for the first-order interaction between X_1 and X_2. Second, any categorical variable is handled via dummy variables. For example, if X has C levels, then $(C-1)$ dummy variables $(Z_1^{(X)}, \ldots, Z_{C-1}^{(X)})$ can be created by setting the last level as baseline, where $Z_C^{(X)} = 1$ if an observation has X in the Cth level, and 0 otherwise. Third, certain nonlinearity in predictors can be integrated by transforming predictor variables. For example, the model form remains linear in $\boldsymbol{\beta}$ after transforming X into its polynomial term of the lth order, X^l.

MODEL ESTIMATION

Model estimation involves estimating the parameters in the model, including both $\boldsymbol{\beta}$ and σ^2. There are several estimation methods available for linear models, including least squares, maximum likelihood, Bayesian approach, robust estimation, ridge regression, and so on. In this article, we shall cover the first two methods only.

Least Squares

The most popular method for estimating $\boldsymbol{\beta}$ is least squares (LS), which minimizes the distance from the observed response to the predicted values,

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{t}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \quad (3)$$

Differentiating with respect to $\boldsymbol{\beta}$ gives

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\left(\mathbf{X}^{\mathrm{t}}\mathbf{y} - \mathbf{X}^{\mathrm{t}}\mathbf{X}\boldsymbol{\beta}\right). \quad (4)$$

Setting Eq. (4) to $\mathbf{0}$ yields the normal equation $\mathbf{X}^{\mathrm{t}}\mathbf{y} = \mathbf{X}^{\mathrm{t}}\mathbf{X}\boldsymbol{\beta}$.

Let us assume that $\mathbf{X}$ is of full column rank p. Thus, the Gram matrix $\mathbf{X}^{\mathrm{t}}\mathbf{X}$ must be positive definite (p.d.). The least squares estimator (LSE) $\hat{\boldsymbol{\beta}}$ exists as a unique solution to the normal equation, and is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{t}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{t}}\mathbf{y}. \quad (5)$$

Subsequently, the vector of fitted values, $\hat{\mathbf{y}}$, is

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^{\mathrm{t}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{t}}\mathbf{y} = \mathbf{H}\mathbf{y}, \quad (6)$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{t}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{t}}$ is often called the hat matrix or the projection matrix.

The LS method has a geometrical representation in $\mathbb{R}^n$. To this end, the response vector $\mathbf{y}$ and all column vectors in $\mathbf{X}$ can be viewed as points in $\mathbb{R}^n$. Let $\mathbb{V}$ be the linear space spanned by the column vectors in $\mathbf{X}$, and denote $\mathbb{V}$ by $C(\mathbf{X}) = \left\{\mathbf{v} \in \mathbb{R}^n : \mathbf{v} = \mathbf{X}\mathbf{b}^* \text{ for some vector } \mathbf{b}^* \in \mathbb{R}^{p+1}\right\}$. Noting that $Q(\boldsymbol{\beta}) = \| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \|^2$ in (3), the LS problem can be restated as minimizing the distance $\| \mathbf{y} - \mathbf{v} \|^2$ subject to $\mathbf{v} \in \mathbb{V}$. In other words, LSE seeks $\mathbf{v} \in \mathbb{V}$ that is closest to $\mathbf{y}$. As illustrated in Figure 1, for a two-dimensional $\mathbb{V}$, the minimum distance can only be achieved by the perpendicular projection $\mathbf{P}_{\mathbb{V}}\mathbf{y}$ of $\mathbf{y}$ onto $\mathbb{V}$. Accordingly, it can be seen easily that $\mathbf{P}_{\mathbb{V}}\mathbf{y} = \mathbf{H}\mathbf{y}$.

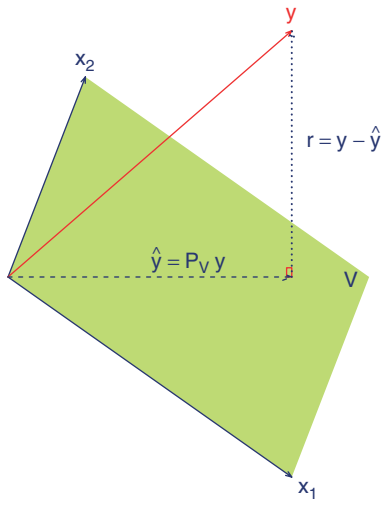


FIGURE 1 | Geometric illustration of least squares estimator.

© 2012 Wiley Periodicals, Inc.

As a by-product of the geometrical approach, the residual vector is readily available

$$\mathbf{r} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{P}_{\mathbb{V}})\mathbf{y} = \mathbf{P}_{\mathbb{V}^\perp}\mathbf{y},$$

where $\mathbf{V}^\perp \subset \mathbb{R}^n$ denotes the subspace perpendicular to $\mathbb{V}$ and $\mathbf{P}_{\mathbb{V}^\perp} = \mathbf{I} - \mathbf{H}$ is its associated projection matrix. It follows that $\mathbf{r} \perp \hat{\mathbf{y}}$, as shown in Figure 1. Moreover, the minimized least squares criterion leads to

$$Q(\hat{\boldsymbol{\beta}}) = \| \mathbf{r} \|^2 = \mathbf{y}^{\mathrm{t}}\mathbf{P}_{\mathbb{V}^\perp}\mathbf{y},$$

which is often referred to as *the residual sum of squares* or *the sum of squares for error* (SSE). Since $E(\mathrm{SSE}) = \sigma^2(n - (p+1))$, a natural unbiased estimator of σ^2 is given by

$$\hat{\sigma}^2 = \mathrm{SSE}/(n - (p+1)). \tag{7}$$

Both $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ enjoy certain optimality properties. The well-known Gauss–Markov theorem states that $\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$, meaning that $\hat{\boldsymbol{\beta}}$ has the minimum variance among all linear unbiased estimators of $\boldsymbol{\beta}$. Note that $\hat{\boldsymbol{\beta}}$ is a vector, and the term 'minimum variance' is used in a general sense. Specifically, if $\check{\boldsymbol{\beta}}$ is an LUE of $\boldsymbol{\beta}$, then $\mathrm{cov}(\check{\boldsymbol{\beta}}) - \mathrm{cov}(\hat{\boldsymbol{\beta}})$ is a nonnegative-definite matrix. This result applies to linear functions of $\boldsymbol{\beta}$ as well. Namely, $\boldsymbol{\Lambda}\hat{\boldsymbol{\beta}}$ is the BLUE of $\boldsymbol{\Lambda}\boldsymbol{\beta}$ with matrix $\boldsymbol{\Lambda}$ being $m \times (p+1)$ of full row rank m. Under some conditions, it can also be shown that $\hat{\sigma}^2$ is the best quadratic unbiased estimator of σ^2.

When $\mathbf{X}$ is not of full column rank, the Gram matrix $(\mathbf{X}^{\mathrm{t}}\mathbf{X})$ is no longer invertible. There are several ways to circumvent the problem, including dropping redundant columns in $\mathbf{X}$, reparameterizing $\boldsymbol{\beta}$, or centering predictors. It is worth noting that centering or standardizing predictors is important not only to achieve numerical stabilities but also to enhance comparability of slope estimates (which is particularly desirable in L_1 regularization), although the interpretability within the application context may be slightly affected. Comparatively, another convenient way of dealing with nonfull rank $\mathbf{X}$ is to employ a generalized inverse $(\mathbf{X}^{\mathrm{t}}\mathbf{X})^-$ in $\mathbf{H}$. With this approach, the LSE $\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{t}}\mathbf{X})^-\mathbf{X}^{\mathrm{t}}\mathbf{y}$ is not uniquely determined unless further constraints are posed. However, the hat matrix $\mathbf{H}$, as well as $\hat{\mathbf{y}}$, remains invariant with different choices of $(\mathbf{X}^{\mathrm{t}}\mathbf{X})^-$. Furthermore, $\boldsymbol{\Lambda}\hat{\boldsymbol{\beta}}$ remains to be the BLUE of $\boldsymbol{\Lambda}\boldsymbol{\beta}$ as long as $\boldsymbol{\Lambda}\boldsymbol{\beta}$ is estimable, meaning that there exists an $m \times n$ matrix $\mathbf{A}$ such that $E(\mathbf{A}\mathbf{y}) = \boldsymbol{\Lambda}\boldsymbol{\beta}$. Throughout the article, we shall assume that $\mathbf{X}$ has a full column rank $(p+1)$.

Maximum Likelihood

Under the normality assumption of the error term in the linear regression model, estimation of $(\boldsymbol{\beta}, \sigma^2)$ can be made via maximum likelihood (ML). With model (2),

$$\mathbf{y}|\mathbf{X} \sim \mathcal{N}\left\{\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}\right\}. \tag{8}$$

For given data, the corresponding likelihood function is

$$\begin{aligned} L(\boldsymbol{\beta}, \sigma^2) &= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{t}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2}\right\} \\ &= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{\boldsymbol{\beta}^{\mathrm{t}}\mathbf{X}^{\mathrm{t}}\mathbf{X}\boldsymbol{\beta}}{2\sigma^2}\right\} \\ &\quad \cdot \exp\left\{\frac{\boldsymbol{\beta}^{\mathrm{t}}\mathbf{X}^{\mathrm{t}}\mathbf{y}}{\sigma^2} - \frac{\mathbf{y}^{\mathrm{t}}\mathbf{y}}{2\sigma^2}\right\} \end{aligned}$$

in the standard exponential family form. It follows that the sufficient and complete statistic for $(\boldsymbol{\beta}^{\mathrm{t}}, \sigma^2)$ is $(\mathbf{y}^{\mathrm{t}}\mathbf{X}, \mathbf{y}^{\mathrm{t}}\mathbf{y})$. The log-likelihood is

$$\begin{aligned} l(\boldsymbol{\beta}, \sigma^2) &= -n/2 \cdot \log(2\pi) - n/2 \cdot \log\sigma^2 \\ &\quad - (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{t}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})/(2\,\sigma^2). \end{aligned} \tag{9}$$

Setting the first derivative of l with respect to $(\boldsymbol{\beta}, \sigma^2)$ to $\mathbf{0}$ yields the maximum likelihood estimator (MLE). The MLE of $\boldsymbol{\beta}$ is exactly the same as its LSE. The MLE of σ^2, $\tilde{\sigma}^2 = \mathrm{SSE}/n$, is biased, although the bias goes to 0 asymptotically.

Following standard ML arguments and the Lehmann and Scheffé[1,2] theorem, $(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2)$ are the (unique) uniformly minimum variance unbiased estimators (UMVUE) of $(\boldsymbol{\beta}, \sigma^2)$, meaning that they have lower variance than any other unbiased estimators for all possible values of $(\boldsymbol{\beta}, \sigma^2)$. Compared to the BLUE concept, UMVUE is not restricted to linear estimators only and hence represents enhanced optimality. However, this property is built upon the additional normality assumption, while Gauss–Markov theorem holds without this assumption.

Besides, it is noteworthy that unbiasedness of estimators is not necessarily an appealing property. In general, the performance of an estimator $\hat{\theta}$ in estimating θ can be measured by its mean squared error (MSE),

$$\mathrm{MSE}(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = \{E(\hat{\theta}) - \theta\}^2 + \mathrm{var}(\hat{\theta}),$$

which is the sum of its squared bias and variance. Shrinkage methods such as ridge regression often provide biased estimators with a larger reduction in variances, resulting in a smaller MSE.

© 2012 Wiley Periodicals, Inc.

STATISTICAL INFERENCE

Statistical inference involves either testing hypotheses or constructing confidence intervals about $(\boldsymbol{\beta}, \sigma^2)$. To facilitate distributional properties, the normal assumption is explicitly made, although asymptotic results can be used for large samples.

Inference on $\boldsymbol{\Lambda\beta}$

We consider the general problem of inferring about $\boldsymbol{\Lambda\beta}$. There are a few different ways to proceed. We first illustrate one method based on multivariate normal distributions in detail and then we discuss other approaches.

First, a natural estimator of $\boldsymbol{\Lambda\beta}$ is $\boldsymbol{\Lambda}\hat{\boldsymbol{\beta}}$. Using Eq. (8),

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{t}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{t}}\mathbf{y} \sim \mathcal{N}_{p+1}\left\{\boldsymbol{\beta}, \sigma^2(\mathbf{X}^{\mathrm{t}}\mathbf{X})^{-1}\right\}$$

and hence

$$\boldsymbol{\Lambda}\hat{\boldsymbol{\beta}} \sim \mathcal{N}_m\left\{\boldsymbol{\Lambda\beta}, \sigma^2\boldsymbol{\Lambda}(\mathbf{X}^{\mathrm{t}}\mathbf{X})^{-1}\boldsymbol{\Lambda}^{\mathrm{t}}\right\}.$$

It follows that

$$\frac{(\boldsymbol{\Lambda}\hat{\boldsymbol{\beta}} - \boldsymbol{\Lambda\beta})^{\mathrm{t}}\left[\boldsymbol{\Lambda}(\mathbf{X}^{\mathrm{t}}\mathbf{X})^{-1}\boldsymbol{\Lambda}^{\mathrm{t}}\right]^{-1}(\boldsymbol{\Lambda}\hat{\boldsymbol{\beta}} - \boldsymbol{\Lambda\beta})}{\sigma^2} \sim \chi^2(m). \tag{10}$$

Second, using the distributional property of a quadratic form of multivariate normal variables, it can be shown that

$$\mathrm{SSE}/\sigma^2 \sim \chi^2(n-(p+1)). \tag{11}$$

Third, $\boldsymbol{\Lambda}\hat{\boldsymbol{\beta}}$ and $SSE = \| \mathbf{r} \|^2$ are independent. To see this, rewrite

$$\begin{aligned}\boldsymbol{\Lambda}\hat{\boldsymbol{\beta}} &= \boldsymbol{\Lambda}(\mathbf{X}^{\mathrm{t}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{t}}\mathbf{y} = \boldsymbol{\Lambda}(\mathbf{X}^{\mathrm{t}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{t}}\{\mathbf{X}(\mathbf{X}^{\mathrm{t}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{t}}\mathbf{y}\} \\ &= \boldsymbol{\Lambda}(\mathbf{X}^{\mathrm{t}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{t}}\hat{\mathbf{y}}.\end{aligned}$$

Namely, $\boldsymbol{\Lambda}\hat{\boldsymbol{\beta}}$ is a linear function of $\hat{\mathbf{y}}$. As $\mathbf{r} \perp \hat{\mathbf{y}}$, their covariance matrix is $0_{n\times n}$. This, together with the fact that they follow a joint multivariate normal distribution, implies independence.

Combining Eqs. (10), (11), and their independence together, it follows that

$$\begin{aligned}F &= \frac{(\boldsymbol{\Lambda}\hat{\boldsymbol{\beta}} - \boldsymbol{\Lambda\beta})^{\mathrm{t}}\left[\boldsymbol{\Lambda}(\mathbf{X}^{\mathrm{t}}\mathbf{X})^{-1}\boldsymbol{\Lambda}^{\mathrm{t}}\right]^{-1}(\boldsymbol{\Lambda}\hat{\boldsymbol{\beta}} - \boldsymbol{\Lambda\beta})/m}{SSE/(n-p-1)} \\ &\sim F^{(m,n-p-1)}\end{aligned} \tag{12}$$

by the definition of the F distribution.

Inference on $\boldsymbol{\Lambda\beta}$ can be made accordingly. Under the null hypothesis, $H_0 : \boldsymbol{\Lambda\beta} = 0$, the observed test statistic is

$$F_{\mathrm{obs}} = \frac{(\boldsymbol{\Lambda}\hat{\boldsymbol{\beta}})^{\mathrm{t}}\left\{\boldsymbol{\Lambda}(\mathbf{X}^{\mathrm{t}}\mathbf{X})^{-1}\boldsymbol{\Lambda}^{\mathrm{t}}\right\}^{-1}(\boldsymbol{\Lambda}\hat{\boldsymbol{\beta}})}{m\hat{\sigma}^2} \overset{H_0}{\sim} F^{(m,n-p-1)}. \tag{13}$$

When $F_{\mathrm{obs}} > F_{1-\alpha}^{(m,n-p-1)}$, we reject the null at significance level α, where $F_{1-\alpha}^{(m,n-p-1)}$ is the upper $1-\alpha$ quantile of the F distribution with m and $n-p-1$ degrees of freedom. Under $H_a : \boldsymbol{\Lambda\beta} = \tilde{\mathbf{b}}$, F_{obs} follows noncentral $F^{(m,n-p-1)}(\delta)$ with a noncentrality parameter $\delta = \tilde{\mathbf{b}}^{\mathrm{t}}\left\{\boldsymbol{\Lambda}(\mathbf{X}^{\mathrm{t}}\mathbf{X})^{-1}\boldsymbol{\Lambda}^{\mathrm{t}}\right\}^{-1}\tilde{\mathbf{b}}/\sigma^2$. In some texts, a factor of $(1/2)$ is added to the definition of δ. Furthermore, an approximate $(1-\alpha)\times 100\%$ confidence set for $\boldsymbol{\Lambda\beta}$ can be obtained from Eq. (12) as the set of $\mathbf{d} \in \mathbb{R}^m$ such that

$$\begin{aligned}\Big\{&(\boldsymbol{\Lambda}\hat{\boldsymbol{\beta}} - \mathbf{d})^{\mathrm{t}}[\boldsymbol{\Lambda}(\mathbf{X}^{\mathrm{t}}\mathbf{X})^{-1}\boldsymbol{\Lambda}^{\mathrm{t}}]^{-1}(\boldsymbol{\Lambda}\hat{\boldsymbol{\beta}} - \mathbf{d}) \\ &\le m\hat{\sigma}^2 \cdot F_{1-\alpha}^{(m,n-p-1)}\Big\}.\end{aligned} \tag{14}$$

Most common inferences in linear models can be viewed as its special cases. For example, the case of $\boldsymbol{\Lambda} = \mathbf{I}$ gives the following $(1-\alpha)\times 100\%$ confidence band for $\boldsymbol{\beta}$

$$\left\{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{t}}\left(\mathbf{X}^{\mathrm{t}}\mathbf{X}\right)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \le (p+1)\cdot\hat{\sigma}^2 \cdot F_{1-\alpha}^{(p+1,n-p-1)}\right\}. \tag{15}$$

The case of $\boldsymbol{\Lambda} = \mathbf{e}_{\mathrm{j}} = (0,\ldots,0,1_{(j)},0,\ldots,0)^{\mathrm{t}}$ corresponds to the inference on an individual parameter β_j. In this case, $m = 1$ and the $F^{(1,n-p-1)}$ distribution, after taking its square root, reduces to the $t^{(n-p-1)}$ distribution.

On the basis of the confidence band in either Eq. (14) or Eq. (15), Scheffé[3] derived simultaneous confidence intervals for all linear combinations of form $\boldsymbol{\tau}^{\mathrm{t}}\boldsymbol{\beta}$ using the Cauchy–Schwarz inequality. Let $\mathbf{a}$ and $\mathbf{b}$ be any vectors with appropriate dimensions and $\mathbf{T}$ be a positive definite symmetric matrix with positive definite symmetric square root $\mathbf{T}^{1/2}$ (i.e., $\mathbf{T} = \mathbf{T}^{1/2}\mathbf{T}^{1/2}$). The Cauchy–Schwarz inequality states that

$$\begin{aligned}&(\mathbf{a}^{\mathrm{t}}\mathbf{b})^2 \le (\mathbf{a}^{\mathrm{t}}\mathbf{a})\cdot(\mathbf{b}^{\mathrm{t}}\mathbf{b}) \\ \Rightarrow\quad &(\mathbf{a}^{\mathrm{t}}\mathbf{b})^2 = \left\{(\mathbf{T}^{1/2}\mathbf{a})^{\mathrm{t}}(\mathbf{T}^{-1/2}\mathbf{b})\right\}^2 \\ &\qquad\le (\mathbf{a}^{\mathrm{t}}\mathbf{T}\mathbf{a})\cdot(\mathbf{b}^{\mathrm{t}}\mathbf{T}^{-1}\mathbf{b}) \\ \Rightarrow\quad &\frac{(\mathbf{a}^{\mathrm{t}}\mathbf{b})^2}{\mathbf{a}^{\mathrm{t}}\mathbf{T}\mathbf{a}} \le \mathbf{b}^{\mathrm{t}}\mathbf{T}^{-1}\mathbf{b}.\end{aligned}$$

 © 2012 Wiley Periodicals, Inc.

Therefore,

$$\mathbf{b}^t\mathbf{T}^{-1}\mathbf{b} \le c_0 \iff \sup_{\mathbf{a}\neq 0}\left\{\frac{(\mathbf{a}^t\mathbf{b})^2}{\mathbf{a}^t\mathbf{T}\mathbf{a}}\right\} \le c_0, \qquad (16)$$

for any positive value c_0. Then, let $\mathbf{b} = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$, $\mathbf{T}^{-1} = \mathbf{X}^t\mathbf{X}$, $c_0 = (p+1)\cdot\hat{\sigma}^2\cdot F_{1-\alpha}^{(p+1,n-p-1)}$ in Eq. (15), and $\boldsymbol{\tau} = a$. Applying Eq. (16), we have that

$$\begin{aligned}\Pr\Big\{&|\,\boldsymbol{\tau}^t\hat{\boldsymbol{\beta}} - \boldsymbol{\tau}^t\boldsymbol{\beta}\,| \\ &\le \sqrt{(p+1)\cdot\hat{\sigma}^2\cdot F_{1-\alpha}^{(p+1,n-p-1)}\cdot\boldsymbol{\tau}^t(\mathbf{X}^t\mathbf{X})^{-1}\boldsymbol{\tau}},\ \forall\,\boldsymbol{\tau}\Big\} \\ &= 1-\alpha. \end{aligned} \qquad (17)$$

Prediction

The prediction problem can be viewed as a special case of the general inference outlined above. There are typically two types of predictions: estimating the mean response $E(y_0)$ or predicting the response value y_0 from a given vector, $\mathbf{x} = \mathbf{x}_0$. To take into account for the intercept, we add 1 to be the first component of $\mathbf{x}_0$.

For the sake of illustrating the difference, we consider an example discussed in the section on *Model Specification* in which SBP is regressed on age and gender. The first scenario is to estimate the average SBP of all males at age 40, while, in the second scenario, we predict an SBP of a person who is male and 40 years old. Clearly, the latter task involves more variability than the former one.

In general, given $\mathbf{x}_0$, inference on $E(y_0) = \mathbf{x}_0^t\boldsymbol{\beta}$ can be obtained by letting $\boldsymbol{\Lambda} = \mathbf{x}_0$ in Eq. (14). After algebraic simplification, a $(1-\alpha)\times 100\%$ confidence interval for mean response $E(y_0)$ is given by

$$\mathbf{x}_0^t\hat{\boldsymbol{\beta}} \pm t_{1-\alpha/2}^{(n-p-1)}\cdot\hat{\sigma}\cdot\sqrt{\mathbf{x}_0^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{x}_0}.$$

To obtain the prediction interval, note that $y_0 = \mathbf{x}_0^t\boldsymbol{\beta} + \varepsilon_0$, where $\varepsilon_0 \sim \mathcal{N}(0,\sigma^2)$ is independent of the sample data that have been used to estimate the linear model. We predict y_0 for the given x_0 by $\hat{y}_0 = \mathbf{x}_0^t\hat{\boldsymbol{\beta}} + \hat{\varepsilon}_0$, where $\hat{\varepsilon}_0 = 0$ as $E(\varepsilon_0) = 0$. However, $\mathrm{var}(\hat{y}_0) = \mathrm{var}(\mathbf{x}_0^t\hat{\boldsymbol{\beta}}) + \mathrm{var}(\hat{\varepsilon}_0) = \sigma^2\cdot(1 + \sqrt{\mathbf{x}_0^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{x}_0})$. The additional unit of σ^2 in its variance is contributed by the extra random error term in the prediction. As a result, a $(1-\alpha)\times 100\%$ prediction interval for $\mathbf{y}_0$ is

$$\mathbf{x}_0^t\hat{\boldsymbol{\beta}} \pm t_{1-\alpha/2}^{(n-p-1)}\cdot\hat{\sigma}\cdot\sqrt{1 + \mathbf{x}_0^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{x}_0}.$$

More generally, simultaneous intervals can be constructed for predictions at multiple $\mathbf{x}_0$ values. One possible approach is to apply the Scheffé method, which was outlined earlier.

Other Approaches for Obtaining the *F* Test

There are other routes to yield the F test in Eq. (13). First of all, let SSE_0 denote the sum of squares for error associated with a restricted or reduced model under H_0,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \text{subject to } \boldsymbol{\Lambda}\boldsymbol{\beta} = 0. \qquad (18)$$

Then, it can be shown that

$$F_{\mathrm{obs}} = \frac{(\mathrm{SSE}_0 - \mathrm{SSE})/m}{\mathrm{SSE}/(n-p-1)}, \qquad (19)$$

where m corresponds to the difference in model complexity (measured by the number of degrees of freedom or the number of parameters) between the full model and the reduced model.

The second route is geometric. Define a subspace $\mathbb{W}\subset\mathbb{V}$:

$$\mathbb{W} = \Big\{\mathbf{w} : \mathbf{w} = \mathbf{X}\mathbf{b}^* \text{ for some } \mathbf{b}^* \in \mathbf{R}^{(p+1)} \text{ that satisfies } \boldsymbol{\Lambda}\mathbf{b} = 0\Big\}. \qquad (20)$$

Then, denote $\mathbb{V}|\mathbb{W}$ as the subspace of all elements in $\mathbb{V}$ that are perpendicular to $\mathbb{W}$. It can be shown that the matrix $\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\boldsymbol{\Lambda}^t$ forms a basis for $\mathbb{V}|\mathbb{W}$. In other words, $\mathbb{V}|\mathbb{W} = C(\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\boldsymbol{\Lambda}^t)$ and $\dim(\mathbb{V}|\mathbb{W}) = m$. After algebraic simplification, we have that

$$F_{\mathrm{obs}} = \frac{\|\,\mathbf{P}_{\mathbb{V}|\mathbb{W}}\,\mathbf{y}\,\|^2/m}{\|\,\mathbf{P}_{\mathbb{V}^\perp}\,\mathbf{y}\,\|^2/(n-p-1)}.$$

This geometric representation is illustrated in Figure 2 when $\dim(\mathbb{V}) = 2$ and $\dim(\mathbb{W}) = 1$.

Finally, the F test also corresponds to the likelihood ratio test in the ML framework. Let $\hat{L}$ denote the maximized likelihood associated with the linear model in Eq. (2). In addition, let $\hat{L}_0$ denote the maximized likelihood associated with the reduced model in Eq. (18), which can be obtained via the Lagrange multiplier technique. Then, it can be shown that

$$F_{\mathrm{obs}} = \frac{n-p-1}{m}\left\{\left(\hat{L}_0/\hat{L}\right)^{-2/n} - 1\right\}.$$

Computer Output of Linear Model Fit

The results obtained from fitting a linear model are often summarized in two tables: the parameter estimates table (see Table 1) and the analysis of variance (ANOVA) table (see Table 2). These two

© 2012 Wiley Periodicals, Inc.

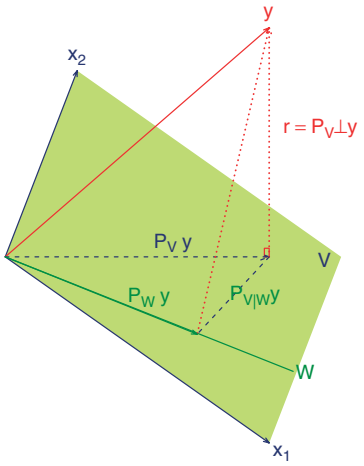


FIGURE 2 | Geometric illustration of the *F* test.

TABLE 1 | Table of Parameter Estimates

Parameter	Estimate	S.E.	t	P-Value
β_0	$\hat{\beta}_0$	$\text{se}(\hat{\beta}_0)$	t_0	$2\,\text{P}\left(t^{(n-p-1)} \geq \lvert t_0 \rvert\right)$
β_1	$\hat{\beta}_1$	$\text{se}(\hat{\beta}_1)$	t_1	$2\,\text{P}\left(t^{(n-p-1)} \geq \lvert t_1 \rvert\right)$
β_2	$\hat{\beta}_2$	$\text{se}(\hat{\beta}_2)$	t_2	$2\,\text{P}\left(t^{(n-p-1)} \geq \lvert t_2 \rvert\right)$

tables are the standard outputs in statistical packages. The parameter estimates table also includes t tests for testing each of the individual regression parameters, $H_0 : \beta_j = 0$ $(j = 1, \ldots, p)$. In contrast, the ANOVA table presents the components involved in a global F test of $H_0 : \beta_1 = \cdots = \beta_p = 0$ for assessing the overall validity of the linear model by comparing the full model versus the null model (no predictors being included). Both t and F tests are special cases of the F test for general linear hypotheses with an appropriate choice of $\mathbf{\Lambda}$.

The ANOVA table also involves a decomposition of the total variation in observed responses, which is directly linked to the F test comparing the full versus reduced models as given in Eq. (19). Specifically, the total variation, measured by the total sum of squares (SST), breaks into two parts, the portion that can be explained by the regression model (measured by the sum of squares for regression or SSR) and the remaining unexplained portion (measured by SSE). Accordingly,

$$\begin{aligned}\text{SST} &= \sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(\overline{y} - \hat{y}_i)^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ &= \text{SSR} + \text{SSE}. \end{aligned} \tag{21}$$

Equation (21) is also called the *fundamental equation of regression*, which holds in general regression problems. The widely used coefficient of determination R^2 is defined as $R^2 = \text{SSR}/\text{SST}$, which can be easily interpreted as the proportion of the total variation in observed responses that can be accounted for by its linear regression on $\mathbf{X}$.

TABLE 2 | The ANOVA Table

Source	DF	SS	MS	Overall F	P-Value
Model	p	SSR	MSR	F_{obs}	$\text{P}\left\{F^{(p,\,n-p-1)} \geq F_{\text{obs}}\right\}$
Error	$n-(p+1)$	SSE	MSE		
Total	$n-1$	SST			

VARIABLE SELECTION

Why Variable Selection?

In the model specifications (1) and (2), we have assumed that the true regression function $\boldsymbol{\mu} = [E(y_i|\mathbf{x}_i)] = [\mu(\mathbf{x}_i)]$ is in the linear form specified by the linearity assumption $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$. This is unlikely to be true in reality, where model misspecification can occur in various ways. For example, the underlying regression function $\mu(\cdot)$ can be curvilinear. Even if it is linear, model specification is still under the risk of overfitting or underfitting or both, meaning that important predictors have been omitted out or irrelevant variables are included in the model. Detailed discussions on variable selection (or model selection) can be found in Linhart and Zucchini,[4] McQuarrie and Tsai,[5] Burnham and Anderson,[6] Claeskens and Hjort,[7] or Konishi and Kitagawa.[8]

To study the adverse effects of underfitting and overfitting on model estimation and prediction, a simplified setting is employed by partitioning the columns of $\mathbf{X}$ into $(\mathbf{X}_1, \mathbf{X}_2)$, where $\mathbf{X}_1$ is $n \times (k+1)$ and $\mathbf{X}_2$ is $n \times (p-k)$. Rewrite model (2) as

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}. \tag{22}$$

At the same time, consider a reduced model that uses $\mathbf{X}_1$ only

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}. \tag{23}$$

Note that we have sightly abused notation by neglecting to distinguish $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$, as well as the error variance σ^2, between the above two models. With this setting, *underfitting* occurs if model (23) is utilized when Eq. (22) is the true model. Let $\hat{\boldsymbol{\beta}}_1$ denote the LSE of $\boldsymbol{\beta}_1$ obtained from fitting Eq. (23). It can be shown that $\hat{\boldsymbol{\beta}}_1$ is biased for $\boldsymbol{\beta}_1$ in model (22), that is, $E(\hat{\boldsymbol{\beta}}_1) \neq \boldsymbol{\beta}_1$, although it has a smaller variance than the LSE of $\boldsymbol{\beta}_1$ obtained from fitting the true model (22).

 © 2012 Wiley Periodicals, Inc.

In contrast, *overfitting* occurs if we fit model (22) when the true model is Eq. (23). In this case, the LSE of $\boldsymbol{\beta}_1$, obtained as a subcomponent of the LSE from fitting Eq. (22), remains unbiased for $\boldsymbol{\beta}_1$ in model (23); however, its variance is inflated when compared to the LSE of $\boldsymbol{\beta}_1$ obtained from fitting the true model (23).

In sum, underfitting leads to bias while overfitting inflates variance. The same conclusion can be drawn from the model predictions. Regression usually has two goals, predicting future observations and studying the relationship between the response and predictors. The latter goal is more related to model interpretation. While one is sometimes more emphasized than the other in specific applications, these two goals are closely related to each other. Reliable interpretation should be based on a model that performs well with new observations. Consider one new observation $(y_0, \mathbf{x}_0)$ first. Let $\hat{y}_0 = \mathbf{x}_0^t \hat{\boldsymbol{\beta}}$ denote its prediction based on a linear model. Then the mean squared error (MSE) of $\hat{y}_0$ is

$$E(\hat{y}_0 - y_0)^2 = E\left\{y_0 - E(\hat{y}_0)\right\}^2 + \text{var}(\hat{y}_0),$$

which is the expected squared bias of $\hat{y}_0$ plus its variance. An overfitted model tends to provide a prediction with a larger variance in spite of a smaller bias, while an underfitted model tends to provide a prediction with a smaller variance yet with a larger bias, a phenomenon often referred to as the 'bias-variance tradeoff'. A reasonably good prediction with a small MSE balances bias and variance.

An empirical illustration of the bias-variance tradeoff is given as follows. Let $\mathcal{D}_0$ denote a new data set consisting of future observations, which are independent of current data $\mathcal{D}$. We fit a number of nested models, including an increasing number of predictors, sequentially ranging from underfitting to overfitting scenarios, and then we compute the resultant sum of squared errors for prediction (SSPE) with $\mathcal{D}_0$. It is important to distinguish between

$$\text{SSPE} = \sum_{i \in \mathcal{D}_0} (y_i - \hat{y}_i)^2 \quad \text{and} \quad \text{SSE} = \sum_{i \in \mathcal{D}} (y_i - \hat{y}_i)^2.$$

Figure 3 plots SSE and SSPE versus model complexity (measured by number of parameters used in the model) based on simulated data. It can be seen that SSE always decreases with additional predictors, even when they have no predictive power. However, SSPE decreases gradually as important variables are added in, hits a minimum near the best model, and then starts to increase when irrelevant variables are included. The graph also suggests that underfitting causes more concern than overfitting if prediction is the primary goal. This is because the inflation amount in SSPE caused by slightly overfitting is relatively smaller than that caused by underfitting. Nevertheless, a simpler model is much easier to interpret.

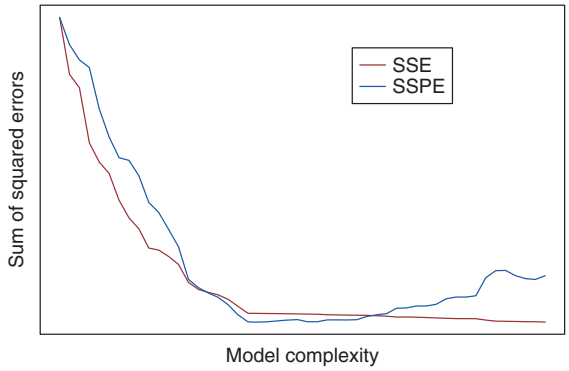


FIGURE 3 | Illustration of the bias-variance tradeoff.

The goal of model selection is to find a parsimonious model that does reasonably well in prediction. There are three approaches for this task, which are discussed in order.

All Possible Regressions

The first approach is *all possible regressions*, which fits all possible subsets of predictors and then selects the best model according to some selection criteria. Note that there are 2^p possible candidate models to be considered. Clearly this method only applies to scenarios where p is small, although there are some methods designed to reduce the computational burden. For example, the *best subsets algorithms* attempt to sort out good model choices while avoiding the evaluation of all models.

Given a model with k predictors and predictor space $\mathbb{V}_0$, a few of the popular model selection criteria for evaluating performance are listed below.

PRESS and GCV

With an additional independent sample, one would naturally consider SSPE as the selection criterion. However, this method is often used for large sample sizes. If the sample size is small, an approximate version of SSPE can be considered via cross-validation. One commonly used criterion, PRESS for *prediction sum of squares*,[9] is obtained via the leave-one-out or jackknife technique, in which each observation is left out in turn and its prediction is computed using the remaining $(n-1)$ observations. As a result,

$$\text{PRESS} = \sum_{i=1}^{n} \left(y_i - \hat{y}_{(-i)}\right)^2, \tag{24}$$

© 2012 Wiley Periodicals, Inc.

where $\hat{y}_{(-i)} = \mathbf{x}_i^t \hat{\boldsymbol{\beta}}_{(-i)}$ denotes the predicted value for y_i by least squares fit on data that leave the ith observation out, and $\hat{\boldsymbol{\beta}}_{(-i)}$ denotes the resulting LSE of $\boldsymbol{\beta}$. Using the fact that

$$\hat{\boldsymbol{\beta}}_{(-i)} = \hat{\boldsymbol{\beta}} - (\mathbf{X}^t\mathbf{X})^{-1} \begin{pmatrix} 1 \\ \mathbf{x}_i \end{pmatrix} \frac{r_i}{1 - h_{ii}},$$

PRESS can be easily computed from the LS fit with the whole data, which is

$$\text{PRESS} = \sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i}{1 - h_{ii}} \right)^2, \tag{25}$$

where h_{ii} is the ith diagonal element of the projection matrix $\mathbf{H}$. Furthermore, replacing the h_{ii}'s by their average, trace($\mathbf{H}$)/n in PRESS/n, Craven and Wahba[10] obtained the generalized cross-validation (GCV) criterion

$$\text{GCV} = \frac{n \cdot \text{SSE}}{\{n - \text{trace}(\mathbf{H})\}^2} = \frac{n \cdot \text{SSE}}{\{n - (k+1)\}^2}. \tag{26}$$

GCV has been extensively used in modern regression methods.

Mallows' C_p

Mallows[11] derived the C_p criterion by examining the expected model error. Given a linear model choice with predictor space $\mathbb{V}_0$, $\text{SSE} = \| \mathbf{y} - \hat{\mathbf{y}} \|^2 = \mathbf{y}^t \mathrm{P}_{\mathbb{V}_0} \mathbf{y}$ provides a measure of the empirical distance between the observed and predicted responses, which leads to

$$E(\text{SSE}) = \boldsymbol{\mu}^t \mathrm{P}_{\mathbb{V}_0} \boldsymbol{\mu} + (n - k - 1)\sigma^2. \tag{27}$$

Using the above result, we obtain the expected model error,

$$\begin{aligned} E \| \boldsymbol{\mu} - \hat{\mathbf{y}} \|^2 &= E \| \boldsymbol{\mu} - \mathrm{P}_{\mathbb{V}_0} \mathbf{y} \|^2 = E \| \boldsymbol{\mu} - \mathrm{P}_{\mathbb{V}_0} (\boldsymbol{\mu} + \boldsymbol{\varepsilon}) \|^2 \\ &= \boldsymbol{\mu}^t \mathrm{P}_{\mathbb{V}_0^\perp} \boldsymbol{\mu} + (k+1)\sigma^2 \\ &= E(\text{SSE}) + \left\{ 2(k+1) - n \right\} \sigma^2. \end{aligned}$$

Accordingly, the C_p criterion is defined as an estimate of $E \| \boldsymbol{\mu} - \hat{\mathbf{y}} \|^2 / \sigma^2$,

$$C_p = \frac{\text{SSE}}{\hat{\sigma}^2} + 2(k+1) - n, \tag{28}$$

where $\hat{\sigma}^2$ is hoped to be a reliable estimate of the true error variance σ^2. In common practice, $\hat{\sigma}^2$ is obtained from the full model that includes all predictors. If a model fits well so that $\boldsymbol{\mu} \in \mathbb{V}_0$ approximately, then $E(C_p) \approx k + 1$. Mallows suggested plotting C_p versus k for all possible models and considering models with $C_p \approx k + 1$ as favorable choices.

AIC and BIC

Akaike[12] derived a criterion from information theories, known as the Akaike information criterion (AIC). As an approximation to the Kullback–Leibler discrepancy function between a candidate model distribution and the true model distribution, AIC is given by

$$\text{AIC} \simeq n \cdot \log(\text{SSE}) + 2 \cdot k$$

up to a constant, which penalizes the goodness-of-fit with model complexity. Later, Hurvich and Tsai[13] proposed an improved Akaike information criterion,

$$\text{AICc} \simeq n \cdot \log(\text{SSE}) + n(n+k)/(n-k-2),$$

which is superior to AIC. In contrast to AIC and AICc, Schwarz[14] employed the Bayesian approach and developed a Bayesian information criterion (BIC), given by

$$\text{BIC} \simeq n \cdot \log(\text{SSE}) + \log(n) \cdot k$$

up to a constant. Since $\log(n) \geq 2$ for $n \geq 8$, BIC imposes a larger penalty for model complexity.

In large samples, a model selection criterion is said to be asymptotically *efficient* if it selects the model with minimum mean squared error, and *consistent* if it selects the true model with probability one. No criterion could be both consistent and efficient. Based on this categorization, PRESS, GCV, C_p, AIC, and AICc are efficient, while BIC is consistent. Detailed illustrations between AIC and BIC criteria can be found in Kuha,[15] Burnham and Anderson,[16] and Yang.[17]

Stepwise Procedure

The second approach is stepwise procedure. It is more feasible for large p, as this selection procedure is designed by adding or removing the predictor variable one at a time. In each step of the procedure, comparison is made only among models that have the same number of variables. Such a comparison can be simply based on SSE. It is noteworthy that dummy variables created for explaining one categorical predictor can be treated as individual variables in the selection process. This essentially involves level merging. Alternatively, this set of dummy variables can be bound together so that we either drop or include them all. In this case, either F test or model selection criterion such as AIC or BIC can be used.

Stepwise procedure can be executed in three ways: backward elimination, forward addition, or

 © 2012 Wiley Periodicals, Inc.

stepwise selection. This procedure mainly uses the F test for model comparison, as implemented in SAS.[18] In backward elimination, one starts with the full model with all predictors being included, and then removes the least significant variable at each step till predictors remaining in the model are all significant. In forward addition, one starts with the null model and adds the most significant variable at each step till no additional variable is significant in the current model. Note that, any predictor that has been removed in backward elimination, has no chance to reenter the model even if its contribution to the current model becomes significant. Similarly, any predictor that has been added in forward addition will not be removed even if its effect becomes insignificant in the current model. Due to these deficiencies, stepwise selection is proposed to amend them. Specifically, it resembles the forward addition, but takes one extra check at each step to remove insignificant variables from the current model. In terms of computational speed, backward elimination is the fastest, followed by forward addition. Yet stepwise selection offers the best performance comparatively.

Despite its popular use in applications, stepwise procedure has been widely recognized as suboptimal in the statistical literature. Because of the multiplicity issue and lack of validation, the selected model is under considerable risk of misidentification and often does not perform well for accommodating new data.

Regularization

The third approach is regularization or shrinkage. In the first two approaches, variable selection is a discrete process, in which a variable is either included or omitted. In contrast, shrinkage methods proceed variable selection in a continuous fashion. Common shrinkage methods optimize the least squares criterion while shrinking the size or length of the regression coefficients. One motivation for this approach is that $E \parallel \hat{\boldsymbol{\beta}} \parallel^2 \geq \parallel \boldsymbol{\beta} \parallel^2$ despite that the LSE, $\hat{\boldsymbol{\beta}}$, is unbiased for $\boldsymbol{\beta}$. In general, a regularized estimator $\tilde{\boldsymbol{\beta}}$ can be obtained as follows:

$$\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \quad \text{subject to} \sum_{j=1}^{p} g(|\beta_j|) \leq t, \tag{29}$$

for some convex (or nonconcave) function $g(\cdot)$ and constant t. Note that the intercept term β_0 can be suppressed by working with centered data. In its equivalent Lagrangian form,

$$\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} g(|\beta_j|) \right\}, \tag{30}$$

where $\lambda > 0$ is a penalty or regularization parameter that controls the amount of the shrinkage.

The power penalty function $g(x) = x^q$ with $q \geq 0$ is often used. The resulting estimator $\tilde{\boldsymbol{\beta}}$ corresponds to the LSE when $q = 0$, the *lasso* (least absolute shrinkage and selection operator[19]) estimator when $q = 1$, and the ridge estimator[20] when $q = 2$. The ridge solution has a simple form

$$\tilde{\boldsymbol{\beta}}_{L_2} = \left(\mathbf{X}^{\mathrm{t}} \mathbf{X} + \lambda \mathbf{I} \right)^{-1} \mathbf{X}^{\mathrm{t}} \mathbf{y}. \tag{31}$$

It can be seen that $\tilde{\boldsymbol{\beta}}_{L_2}$ is biased for $\boldsymbol{\beta}$. However, its MSE can be smaller than that of $\hat{\boldsymbol{\beta}}$ with appropriate choice of λ. Furthermore, the ridge estimator can be minimax under some conditions.[21] While the lasso estimator $\tilde{\boldsymbol{\beta}}_{L_1}$ does not have an explicit form, its entire solution path for any λ can be efficiently obtained via the LARS[22] algorithm. In addition to the power penalty function, Fan and Li[23] proposed the smoothly clipped absolute deviation (SCAD) penalty and established the oracle properties of SCAD estimators. It is of interest to note that the GCV[19] and BIC[24] criteria are often used to determine the optimal λ in ridge regression, lasso, and SCAD.

Both ridge regression and lasso usually provide competitive predictive performance. However, there is a critical difference between the ridge and the lasso estimators. As λ increases, lasso, and its variants, effectively proceed variable selection by setting some coefficients to zero. To gain insight, first observe that

$$\begin{aligned} \parallel \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \parallel^2 &= \parallel \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta} \parallel^2 \\ &= \parallel \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} \parallel^2 + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^{\mathrm{t}} \mathbf{X}^{\mathrm{t}} \mathbf{X} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}), \end{aligned}$$

where the first term does not involve $\boldsymbol{\beta}$. Thus, we are able to rewrite the optimization problem in Eq. (29) as

$$\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^{\mathrm{t}} \mathbf{X}^{\mathrm{t}} \mathbf{X} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \quad \text{subject to} \sum_{j=1}^{p} |\beta_j|^q \leq t. \tag{32}$$

The objective function is a hyper ellipsoid centered at the LSE, $\hat{\boldsymbol{\beta}}$, while the constraint is a disk when $q = 2$ and a diamond when $q = 1$. A graphical

 © 2012 Wiley Periodicals, Inc.

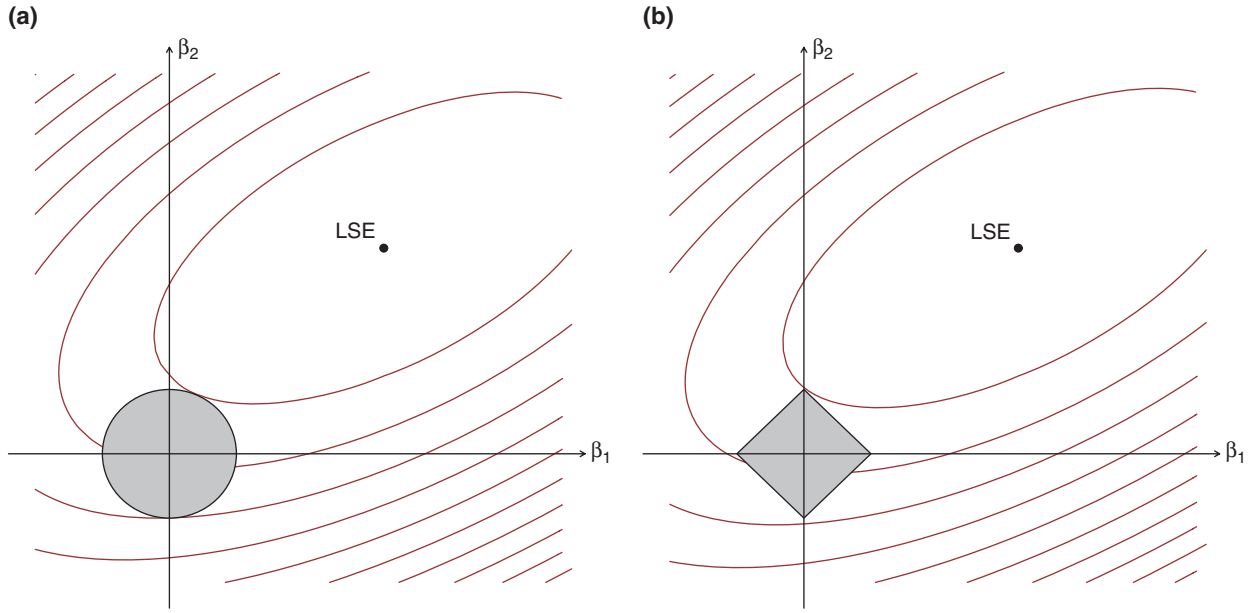


FIGURE 4 | Illustration of shrinkage estimators in the two-dimensional case: (a) ridge ($q = 2$) and (b) lasso ($q = 1$).

illustration for the two dimensional case is given in Figure 4. For both ridge and lasso, the solution is where the elliptical contours touch the boundary of the constraint region. In contrast to the disk case, the diamond constraint region has corners and is very likely to have solution at a corner. When this happens, one parameter estimate becomes zero.

The lasso method has been shown quite successful in both predictive modeling and variable selection. Since the seminal work of Tibshirani,[19] intensive research effort has been devoted to this direction. Lasso variants have been developed and shown to be consistent in both variable selection and estimation; a useful reference can be found in Tibshirani.[25] Analogously, SCAD has also been widely used in selecting variables and estimating regression coefficients simultaneously. Moreover, both Lasso and SCAD have attracted attention in ultrahigh dimensional data analysis.[26–28]

MODEL DIAGNOSTICS

Once a 'best' model is selected, the next step is model diagnostics, which involves three specific tasks: checking model assumptions, detecting outliers, and evaluating computational problems. We discuss each of them given below.

Assumption Checking

As explained, the four major assumptions are all posed on the error terms. Thus it is natural to check assumptions via analysis of the residuals, which can be viewed as the empirical realizations of the error terms. There are several types of residuals, which are listed below in an ascending order of preference.

1. The raw residual $r_i = y_i - \hat{y}_i$ mimics the error term $\varepsilon_i = y_i - \mu_i$.
2. Motivated by the fact that $\varepsilon_i/\sigma \sim \mathcal{N}(0, 1)$, the standardized residual is defined as $z_i = r_i/\hat{\sigma}$.
3. Noting that $\text{var}(r_i) = \sigma^2(1 - h_{ii})$, the studentized residual is defined as $t_i = r_i/\sqrt{\hat{\sigma}^2(1 - h_{ii})}$. If the model is true, then $t_i \sim \mathcal{N}(0, 1)$ approximately.
4. In order to achieve independence between y_i and its predicted value, the prediction of y_i is calculated from the data by omitting the ith observation; the same idea is used in obtaining PRESS. The deleted residual is defined as $e_{(-i)} = y_i - \hat{y}_{(-i)} = r_i/(1 - h_{ii})$, where the definition of $\hat{y}_{(-i)}$ has been introduced in Eqs. (24) and (25).
5. Finally, the studentized deleted residual (also called the jackknife residual), given by

$$r_{(-i)} = \frac{r_i}{\sqrt{\sigma^2_{(-i)}(1 - h_{ii})}} = t_i \sqrt{\frac{n - p - 2}{n - p - 1 - t_i^2}}, \tag{33}$$

where $\hat{\sigma}^2_{(-i)}$, the estimate of σ^2 based on the sample without the ith observation, can

 © 2012 Wiley Periodicals, Inc.

be computed via $(n-p-2)\,\hat{\sigma}^2_{(-i)} = (n-p-1)\,\hat{\sigma}^2 - r_i^2/(1-h_{ii})$.

The jackknife residual $r_{(-i)}$ is the most preferable residual for model diagnoses. Since $\hat{\sigma}^2_{(-i)}$ in Eq. (33) is independent of $\hat{\boldsymbol{\beta}}_{(-i)}$ and hence $r_{(-i)}$, it can be verified that $r_{(-i)} \sim t^{(n-p-2)}$ exactly if the model assumptions are correct. Moreover, the jackknife residuals can be easily computed using the second formula in Eq. (33).

It is a common practice to plot $r_{(-i)}$ versus the predicted values $\hat{y}_i$. As $\mathbf{r} \perp \hat{\mathbf{y}}$, $r_{(-i)}$ and $\hat{y}_i$ are independent of each other. If the model assumptions are valid, the jackknife residuals are expected to randomly scatter around the horizontal line $y = 0$, as shown in Figure 5(a). On the other hand, any systematic nonrandom pattern of the jackknife residuals may indicate some violation of the assumptions in one way or another. Also superimposed on Figure 5(a) are two reference lines from the 2.5th and 97.5th percentiles of $t^{(n-p-2)}$, which can be tentatively used for outlier identification in the spirit of Fisher's least significance difference (LSD) method.

Normality

Note that $r_{(-i)} \sim t^{(n-p-2)}$, which is approximated by $\mathcal{N}(0, 1)$ when $n \gg p$. The informal histogram or quantile–quantile (Q–Q) plot of $r_{(-i)}$'s can be used to examine the normality assumption. In addition, various goodness-of-fit formal tests, such as the Pearson's χ^2 test, Shapiro–Wilk[29] test, or Kolmogorov–Smirnov test, have been used to formally test for normality.

Independence

Examining the assumption of independence among errors (or response observations) is not an easy task. There are only a few limited tests available. However, the plausibility of independence usually can be inspected from the experiment design or the way the data are collected. One common violation of independence occurs when observations are taken as a sequence in order of time and hence exhibit serial correlation. Graphically, the plot of $r_{(-i)}$ versus the sequence order i (or the lag plot of residuals) can be used to examine the dependence of errors. Furthermore, the run tests[30] can provide a rough check for randomness. Moreover, the Durbin–Watson[31,32] statistic and the autocorrelation function (ACF) test can be used to detect autocorrelation.

Homoscedasticity

The assumption of homoscedasticity or equal variances can be inspected from the residual plot. For example, Figure 5(b) illustrates one scenario typically encountered with financial price data, where the error variance increases with the predicted value. It is interesting to note that the LSE remains unbiased under unequal error variances but is no longer BLUE. Formal tests for constant error variances include the White's[33] test, Cook and Weisberg's[34] score test, and several others, all checking whether the variability in e_i or e_i^2 can be accounted for by regressing it on

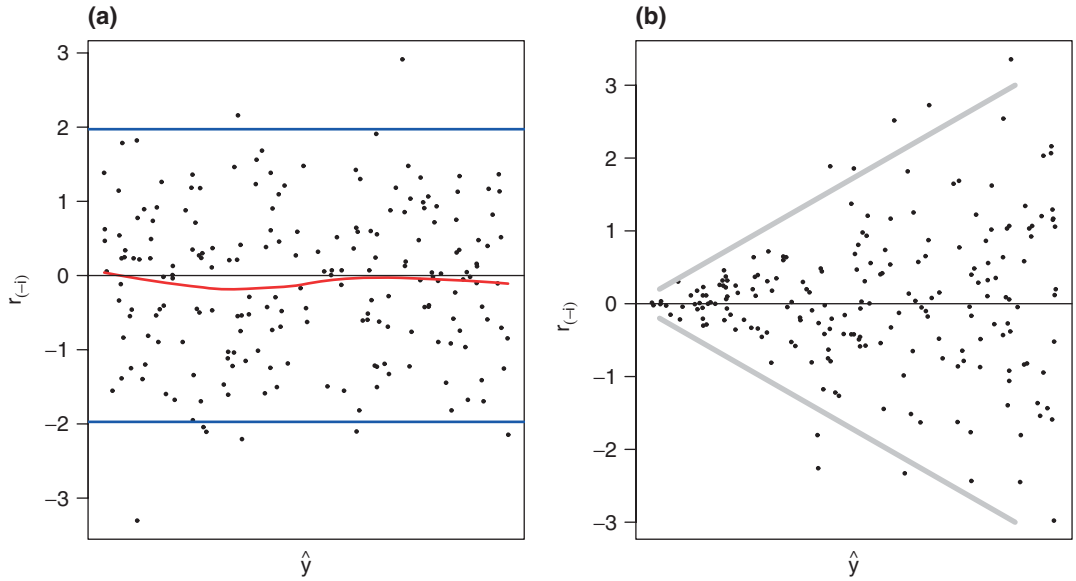


FIGURE 5 | Plot of jackknife residuals $r_{(-i)}$ vs. $\hat{y}_i$: (a) the case where all the model assumptions are valid, superimposed with a smooth curve from loess smoothing; and (b) the case with unequal error variances.

© 2012 Wiley Periodicals, Inc.

the predictors **X** (or the estimate of mean response, $\hat{y}$). Another natural approach is to incorporate the error variance function explicitly in the model setting, and then check whether it reduces to constant variance.

Linearity

Inadequacy of linearity (i.e., linear in regression parameters) can be a serious problem. While the residual plot provides useful diagnostic information for this problem, it does not generally supply any clues as to the true functional form. Toward this end, partial residual plots have been recommended. The ith partial residual for X_j is defined as

$$r_i^{(j)} = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_{j-1} x_{i(j-1)} + \hat{\beta}_{j+1} x_{i(j+1)} + \cdots + \hat{\beta}_p x_{ip}\right)$$

or $r_i^{(j)} = r_i + \hat{\beta}_j x_{ij}$ for $j = 1, \ldots, p$. The plot of $r_i^{(j)}$ versus x_{ij} provides a pictorial exploration of the appropriate functional form for one individual predictor X_j after including other predictors. Figure 6 gives three examples that reflect different diagnostic interpretations regarding the functional form of X_j: (a) X_j might not be needed from the current model; (b) X_j should be included in linear form; (c) A curvilinear form of X_j is needed. Another similar tool, the partial leverage regression plot (i.e., the added variable plot), plots the residuals from the linear model that regresses Y on predictors without X_j against the residuals from the linear model that regresses X_j on other predictors, and this plot can be interpreted in the same manner as the partial residual plot.

Outlier Detection

The second task of model diagnostics is to detect or identify outlying observations. From the perspective of sensitivity analysis, variable selection is concerned about the influence of each column in **X** on model estimation while outlier detection is concerned about the influence of each row of the data. In the regression setting, an observation or row in **X** could be outlier mainly in three ways: outlier in **x**-space; outlier in y-space; or being an influential point that affects the estimation of $\hat{\boldsymbol{\beta}}$ and model prediction. It is noteworthy that the local influence measure[35,36] can be used to assess the effect of minor perturbations of the data, which supplements conventional outlier detections.

Outlier in **x***-Space*

An observation is said to have high leverage if it is outlier in terms of its predictor $\mathbf{x}_i$ value. This can be assessed by the leverage h_{ii}, which is closely related to the Mahalanobis distance from each $\mathbf{x}_i$ to the center $\overline{\mathbf{x}} = \sum_{i=1}^n \mathbf{x}_i/n$. Let $S_X = \sum_{i=1}^n (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^{\mathrm{t}}/(n-1)$ denote the variance–covariance matrix of $\mathbf{x}_i$. Then, the Mahalanobis distance is $d_i = \sqrt{(\mathbf{x}_i - \overline{\mathbf{x}})^{\mathrm{t}} S^{-1} (\mathbf{x}_i - \overline{\mathbf{x}})}$. It can be shown that $h_{ii} = 1/n + (n-1) \cdot d_i^2$. Thus, an observation with high leverage is the one that is distant from the center of points in the **x**-space. The value of h_{ii} ranges from $1/n$ to 1 with average $(p+1)/n$. Points with $h_{ii} > 2(p+1)/n$ are often considered outliers in **x**-space.

Outlier in y*-Space*

A response observation y_i is identified to be an outlier if the observation is sufficiently different from its predicted value. The jackknife residual $r_{(-i)}$ is recommended for this assessment. Since $r_{(-i)} \sim t^{(n-p-2)}$, the

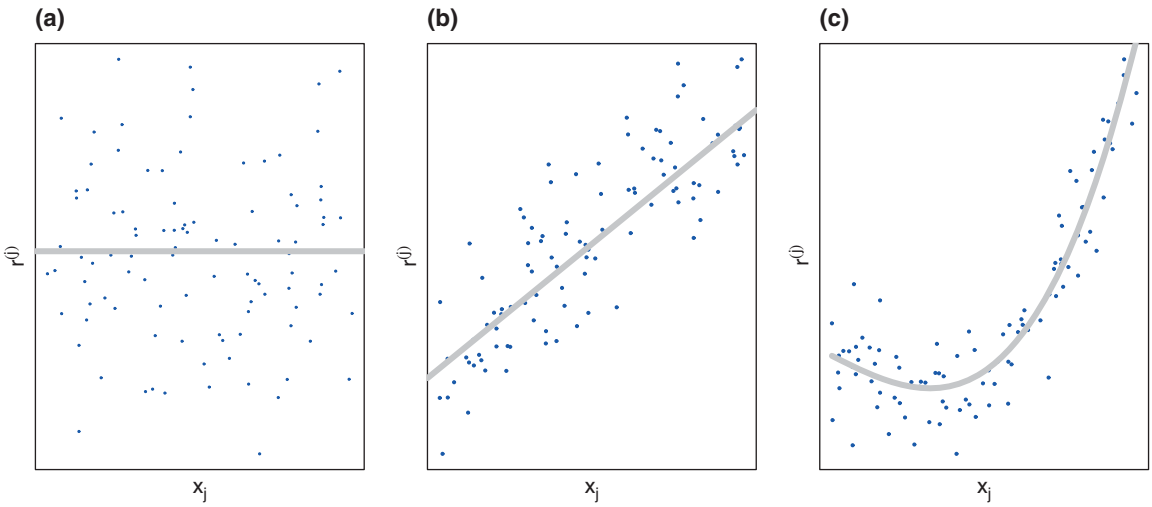


FIGURE 6 | Partial residual plots.

 © 2012 Wiley Periodicals, Inc.

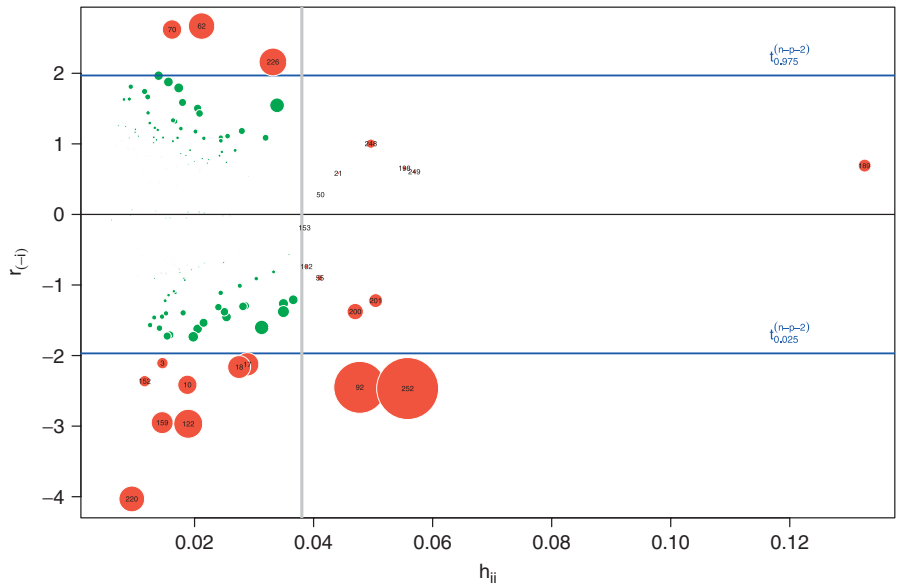


FIGURE 7 | Diagnostic plot of $r_{(-i)}$ vs. h_{ii} for the 1987 baseball salary data. The size of the bubble corresponds to Cook's distance D_i.

2.5th and 97.5th percentiles from $t^{(n-p-2)}$ may be used as benchmarks, yet at the risk of multiplicity.

Influential Points

An observation is said to be an influential point if its removal or inclusion causes dramatic change in model estimations or predictions. The delete-one jackknife technique is the natural approach to tackle this issue. There are many measures developed depending on the specific aspect to be examined. First, DFBETA examines the influence of each observation on each $\hat{\beta}_j$,

$$\text{DFBETA}_{ij} = \frac{\hat{\beta}_j - \hat{\beta}_{j(-i)}}{\sqrt{\hat{\sigma}^2_{(-i)} \cdot (\mathbf{X}^t\mathbf{X})^{-1}_{jj}}},$$

where $\hat{\beta}_{j(-i)}$ denotes the LSE of β_j without the ith observation and $(\mathbf{X}^t\mathbf{X})^{-1}_{jj}$ is the jth diagonal element of matrix $(\mathbf{X}^t\mathbf{X})^{-1}$. Second, DFFITS examines the influence of each observation on its own fitted value,

$$\text{DFFITS}_{ij} = \frac{\hat{y}_i - \hat{y}_{(-i)}}{\sqrt{\hat{\sigma}^2_{(-i)} \cdot h_{ii}}} = r_{(-i)} \cdot \sqrt{\frac{h_{ii}}{1-h_{ii}}}.$$

The ultimate measure for detecting influential points is Cook's distance,[37]

$$D_i = \frac{\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)}\right)^t \mathbf{X}^t\mathbf{X} \left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)}\right)}{(p+1) \cdot \hat{\sigma}^2}$$

$$= \frac{\| \hat{\mathbf{y}} - \hat{\mathbf{y}}_{(-i)} \|^2}{(p+1) \cdot \hat{\sigma}^2} = \frac{r^2_{(-i)}}{p+1} \cdot \frac{h_{ii}}{1-h_{ii}}.$$

Muller and Mok[38] studied the distribution of D_i and provided some critical values. However, the multiplicity issue remains a concern when using these critical values for outlier detection in practice. For the sake of simplicity, one may use the benchmark of 1 to help identify potential outliers.[39]

As a short illustration, we consider the 1987 baseball salary data originally from the 1988 ASA exposition competition. Hoaglin and Velleman[40] found that the following model provides a good fit to the data.

$$\begin{aligned}\log(\text{salary}) = \beta_0 &+ \beta_1 \frac{\text{runcr}}{\text{yrs}} + \beta_2 \sqrt{\text{run86}} \\ &+ \beta_3 \min\{(\text{yrs}-2)_+, 5\} \\ &+ \beta_4 (\text{yrs}-7)_+ + \varepsilon. \end{aligned} \quad (34)$$

Here, function $(x)_+ = x$ if $x > 0$, and 0 otherwise. We refer interested readers to the work of Hoaglin and Velleman[40] for a detailed description of the data set and analysis. The model in Eq. (34) remains linear with transformed variables. On the basis of this model with $n = 263$ and $p = 4$, Figure 7 provides a bubble plot of the three diagnostic measures, $r_{(-i)}$, h_{ii}, and D_i. Twenty-four potential outliers are found: 11 outliers are in x-space detected via the benchmark $2(p+1)/n = 0.038$, 11 outliers are in y-space identified by the benchmarks $t(0.025, n-p-2) = -1.969$ and $t(0.975, n-p-2) = 1.969$, and two outliers (observations 92 and 252) are in both x-space and y-Space. In addition, the Cook's distance measure indicate that the observation

© 2012 Wiley Periodicals, Inc.

252 has a large influence on regression parameter estimates, determined by either the benchmark 1 or Muller and Mok's critical value.

Multicollinearity

The third task in model diagnostics is to detect computational problems in the model fit. One common issue is multicollinearity or collinearity. Multicollinearity occurs when two or more predictors in the linear model are highly correlated with each other. When this is the case, the matrix $\mathbf{X}$ is of nonfull rank and the Gram matrix $\mathbf{X}^t\mathbf{X}$ is singular. Recall that the inverse of $\mathbf{X}^t\mathbf{X}$ is needed in obtaining both $\hat{\boldsymbol{\beta}}$ and its variance–covariance matrix. Thus, the first method for detecting multicollinearity is to consider the spectral decomposition of $\mathbf{X}^t\mathbf{X}$. Let $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_p$ denote the eigenvalues of $\mathbf{X}^t\mathbf{X}$. If $\mathbf{X}^t\mathbf{X}$ is not positive definite, some of its eigenvalues are zero. If the condition number, defined as $\sqrt{\lambda_1/\lambda_p}$, is very large, then multicollinearity could be present.

When multicollinearity occurs, the standard errors (SE) of some $\hat{\beta}_j$'s can be unreasonably large. To see why, a closer look reveals that

$$\mathrm{SE}(\hat{\beta}_j) = \frac{s_y}{s_j}\sqrt{\frac{1-R^2_{Y|\mathbf{X}}}{\left(1-R^2_{X_j|\mathbf{X}_{(-j)}}\right)\cdot(n-p-1)}}, \qquad (35)$$

where s_y and s_j are the sample standard deviation of y and x_j, respectively; $R^2_{Y|\mathbf{X}}$ denotes the R^2 obtained by regressing Y on $\mathbf{X}$; and $R^2_{X_j|\mathbf{X}_{(-j)}}$ denotes the resulting R^2 value from regressing the jth predictor X_j on the remaining predictors $\mathbf{X}_{(-j)}$. If X_j can be expressed as a linear combination of other predictors, $R^2_{X_j|\mathbf{X}_{(-j)}}$ would be 1 and $\mathrm{SE}(\hat{\beta}_j)$ in (35) is infinite. Accordingly, the second measure for detecting collinearity is through the variance inflation factor (VIF), defined by

$$\mathrm{VIF}_j = \frac{1}{1-R^2_{X_j|\mathbf{X}_{(-j)}}}$$

for $j = 1, \ldots, p$. The name of VIF comes from the following observation. Suppose that we are working with normalized or standardized data, in which case $\mathbf{X}^t\mathbf{X}$ becomes the correlation matrix $\mathbf{R}_\mathbf{X}$ among predictors. From $\mathrm{cov}(\hat{\boldsymbol{\beta}}) = \sigma^2\mathbf{R}_\mathbf{X}^{-1}$, it can be found that

$$\mathrm{var}(\hat{\beta}_j) = \sigma^2 \cdot \mathrm{VIF}_j.$$

If the columns in $\mathbf{X}$ are independent, then $\mathbf{R}_\mathbf{X} = \mathbf{I}$ and hence $\mathrm{var}(\hat{\beta}_j) = \sigma^2$. Therefore, VIF_j shows how much $\mathrm{var}(\hat{\beta}_j)$ is inflated by the multicollinearity between X_j and the remaining predictors in $\mathbf{X}$, when compared to the independent case. In practice, a maximum of VIF in excess of 10 is often considered as an indication of multicollinearity. Multicollinearity results in an ill-conditioned Gram matrix $\mathbf{X}^t\mathbf{X}$. To proceed with LS estimation, several handling methods are common, including the removal of redundant predictors, the use of centering data, the generalized inverse, or the ridge regression as discussed earlier.

REMEDIAL MEASURES FOR MODEL REFINEMENT

Various remedial measures are available for refining models and dealing with the problems identified in model diagnostics. This section briefly summarizes some of these techniques.

Variable Transformation

Variable transformation has been widely studied and can be very helpful in improving linearity, stabilizing error variance, and improving normality, although it complicates the model interpretability in practice.

Box–Cox Transformation

In general, power transformation is applicable for predictors,[41] response,[42] and both.[43] A useful reference can be found in the work of Atkinson.[44] For the sake of simplicity, we illustrate only the Box–Cox transformation on the response. To this end, we assume that there is a power transformation such that the transformed model $y_i^{(\lambda)} = \beta_0 + \sum_{j=1}^p \beta_j x_{ij} + \varepsilon_i$ fits well to the data. The family of power transformations from y to $y^{(\lambda)}$ is given by

$$y^{(\lambda)} = \begin{cases} (y^\lambda - 1)/\lambda & \text{if } \lambda \neq 0, \\ \log y & \text{if } \lambda = 0. \end{cases} \qquad (36)$$

To estimate the transformed model, the maximum likelihood approach is used. In particular, a profile likelihood for λ is obtained by substituting $(\boldsymbol{\beta}, \sigma^2)$ with their MLEs. As a result, the MLE of λ can be found. Furthermore, the likelihood ratio test (LRT) of $H_0 : \lambda = 1$ reveals the adequacy of linearity. In addition, a confidence interval for λ can be constructed by inverting the LRT. Bickel and Doksum[45] noticed that the variance of $\hat{\boldsymbol{\beta}}$ associated with transformed model is inflated, relative to the estimate obtained with known λ. For prediction purposes, Carroll and Ruppert[46] found that the prediction $\hat{y}$ obtained by transforming $\hat{y}^{(\lambda)}$ back to its original scale does not have such a problem. Box and Cox[47] suggested employing their method to estimate λ, and then estimating $\boldsymbol{\beta}$ by treating $\hat{\lambda}$ as fixed.

 © 2012 Wiley Periodicals, Inc.

Variance-Stabilizing Transformation

When the nonconstant error variance is a function of the mean such that $\text{var}(y_i) = w(\mu_i)$, (e.g., see Figure (5)(b)), there is a special transformation that helps stabilize the variance. As

$$\text{var}\{f(y)\} \approx \left(\frac{df}{d\mu}\right)^2 \text{var}(y) = \left(\frac{df}{d\mu}\right)^2 \cdot w(\mu),$$

the transformed responses $f(y_i)$ would have approximately constant variance if $f(\cdot)$ is chosen as

$$f(\mu) = \int \frac{d\mu}{\sqrt{w(\mu)}}. \tag{37}$$

Additive Models

The additive model[48] provides a flexible nonparametric way of exploring the functional forms of predictors. Its general form is given by

$$y_i = \beta_0 + \sum_{j=1}^{p} f_j(x_{ij}) + \varepsilon_i, \tag{38}$$

where $f_j(\cdot)$ is an unknown smooth function of x_j. To estimate the $f_j(\cdot)$, an iterative backfitting algorithm iterates between computing the partial residual

$$e_{ij} = y_i - \left\{\hat{\beta}_0 + \sum_{j' \neq j} \hat{f}_{j'}(x_{ij'})\right\},$$

and updating $\hat{f}_j(\cdot)$ as the scatterplot smoother that regresses e_{ij} on x_{ij}. Hastie and Tibshirani[49] proposed generalized additive models (GAM) by extending this idea to handle other types of responses.

ACE and AVAS

Along the same lines as additive models, Breiman and Friedman[50] proposed the alternating conditional expectation (ACE) algorithm to find nonparametric optimal transformations for response and predictors. The working model of ACE is

$$g(y_i) = \beta_0 + \sum_{j=1}^{p} h_j(x_{ij}) + \varepsilon_i. \tag{39}$$

To motivate ACE, consider the minimization of the squared-error loss $E\left\{g(Y) - h(X)\right\}^2$ in the case of $p = 1$. For fixed g, $h^\star(X) = E\{g(Y)|X\}$ minimizes the loss; conversely, for fixed h, $g^\star(Y) = E\{h(X)|Y\}$ minimizes the loss. The key idea of ACE is to alternate between computation of these two conditional expectations. ACE treats Y and X_j's symmetrically, but its performance is unstable. To remedy this problem, Tibshirani[51] proposed the AVAS (additivity and variance stabilization) algorithm by incorporating an additional step with Eq. (37).

Generalized Least Squares

Violation of either the independence or the homoscedasticity assumption on the regression errors has direct effects on parameter estimations and inferences. In contrast to the model given in Eq. (2), we consider a more general model setting,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad \text{with } \boldsymbol{\varepsilon} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 \cdot \boldsymbol{\Sigma}\right). \tag{40}$$

Here, matrix $\boldsymbol{\Sigma}$ is $n \times n$ positive definite with a known form up to several parameters. Accordingly, both $\boldsymbol{\Sigma}^{1/2}$ and $\boldsymbol{\Sigma}^{-1/2}$ exist such that $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{1/2}$ and $\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{-1/2} = \mathbf{I}$.

Left-multiplying $\boldsymbol{\Sigma}^{-1/2}$ on both sides of Eq. (40) yields

$$\boldsymbol{\Sigma}^{-1/2}\mathbf{y} = \boldsymbol{\Sigma}^{-1/2}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\varepsilon},$$

where $\text{cov}(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\varepsilon}) = \sigma^2 \cdot \mathbf{I}$. Therefore, letting

$$\mathbf{y}_0 = \boldsymbol{\Sigma}^{-1/2}\mathbf{y}, \ \mathbf{X}_0 = \boldsymbol{\Sigma}^{-1/2}\mathbf{X}, \text{ and } \mathbf{e} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\varepsilon}, \tag{41}$$

model (40) becomes an ordinary linear model

$$\mathbf{y}_0 = \mathbf{X}_0\boldsymbol{\beta} + \mathbf{e} \quad \text{with } \mathbf{e} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 \cdot \mathbf{I}\right). \tag{42}$$

The estimations and inferences with model (40) can be processed via model (42) and then re-expressed with the quantities in Eq. (41). For example, the least squares criterion becomes

$$\begin{aligned} Q(\boldsymbol{\beta}) &= (\mathbf{y}_0 - \mathbf{X}_0\boldsymbol{\beta})^{\text{t}}(\mathbf{y}_0 - \mathbf{X}_0\boldsymbol{\beta}) \\ &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\text{t}}\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \end{aligned} \tag{43}$$

which is termed the *generalized least squares* (GLS) criterion. The resulting GLS estimator of $\boldsymbol{\beta}$ in model (40) is simply the ordinary LSE in model (42), and it is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}_0^{\text{t}}\mathbf{X}_0)^{-1}\mathbf{X}_0^{\text{t}}\mathbf{y}_0 = \left(\mathbf{X}^{\text{t}}\boldsymbol{\Sigma}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^{\text{t}}\boldsymbol{\Sigma}^{-1}\mathbf{y}. \tag{44}$$

with

$$\text{cov}(\hat{\boldsymbol{\beta}}) = \sigma^2 \cdot \left(\mathbf{X}_0^{\text{t}}\mathbf{X}_0\right)^{-1} = \sigma^2 \cdot \left(\mathbf{X}^{\text{t}}\boldsymbol{\Sigma}^{-1}\mathbf{X}\right)^{-1}.$$

Heteroscedasticity, where observations remain independent but have unequal error variances, can be viewed as a special case of GLS. In this case,

© 2012 Wiley Periodicals, Inc.

$\boldsymbol{\Sigma} = \text{diag}(w_i)$ and the GLS criterion in (43) becomes

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{1}{w_i} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 .$$

The resulting estimate of $\boldsymbol{\beta}$ is termed the *weighted least squares* (WLS) estimate. The WLS technique is critical in nonlinear regression, generalized linear models, L_p estimation, and many other estimation settings.

Another important special case of GLS is when the random errors follow an autoregressive process. For example, in the first-order AR(1) model, $\varepsilon_i = \rho\, \varepsilon_{i-1} + \nu_i$ with $\nu_i \sim \mathbf{N}(0, \sigma^2)$ and $-1 \le \rho \le 1$. The Durbin–Watson[32] test is derived from this approach.

GLS is the primary approach to clustered or longitudinal data, where the complex dependence structure is explicitly formulated via random-effects or mixed-effects models. For example, consider a linear mixed model[52]

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\gamma} \sim \mathcal{N}(\mathbf{0}, \mathbf{W}_1)$ and $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{W}_0)$ are independent, and matrix $\mathbf{Z}$ contains the cluster or time variables that induce dependence. It follows that $\boldsymbol{\Sigma} = \mathbf{Z}\mathbf{W}_1\mathbf{Z}^t + \mathbf{W}_0$, and GLS can be applied. Detailed discussions on parameter estimators and statistical inference can be found in Demidenko.[53]

Robust Regression

Robust regression enables to produce parameter estimators that are less affected by outliers. Hence, it is a useful approach of handling outliers. A known example is Huber's[54] robust regression, which optimizes a criterion of form

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} \phi_a \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right),$$

where

$$\phi_a(r) = \begin{cases} r^2/2 & \text{if } |r| \le a, \\ a|r| - a^2/2 & \text{if } |r| > a, \end{cases} \qquad (45)$$

for some constant a. Minimizing Huber's function leads to a quadratic programming problem.

Many other forms of $\phi(\cdot)$ are available in the literature.[55] One option is taking $\phi(\cdot) = |\cdot|$, which yields the *least absolute deviations* (LAD) estimator. Specifically, the LAD estimator of $\boldsymbol{\beta}$ seeks to minimize the sum of the absolute values of the residuals

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left| y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right| .$$

This robust estimator is usually solved via the linear programming approach, but might have multiple solutions. Another useful method of handling outliers is via WLS so that the smaller weights are assigned to potential outliers.

COMPUTATION IN LINEAR REGRESSION

Computation in linear regression involves algorithms that are used to compute various quantities including $\hat{\boldsymbol{\beta}}$, SSE, $\text{cov}(\hat{\beta})$, the F test statistic, $\mathbf{H}$, etc., and execute different operations such as adding or removing variables. There are three basic methods: the LU or Cholesky decomposition of the Gram matrix $\mathbf{X}^t\mathbf{X}$; QR decomposition of $\mathbf{X}$; and the singular value decomposition (SVD) of $\mathbf{X}$. They are closely related to each other. Among these three, QR decomposition is commonly used in software implementation. The SVD method is most numerically stable, but also most computationally expensive. Interested readers are referred to either Gentle[56] or Seber and Lee[57] for details.

Cholesky Decomposition of $\mathbf{X}^t\mathbf{X}$

The first method aims to solve the normal equation $\mathbf{X}^t\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^t\mathbf{y}$ directly via p Gaussian elimination (GE) steps. Each GE step can be expressed in matrix form as premultiplication by a lower triangular matrix with unit diagonal. The resulting matrix $\mathbf{L}_0$ that carries out the whole GE procedure is also a lower triangular matrix with unit diagonal. GE essentially transforms $\mathbf{X}^t\mathbf{X}$ into an upper-triangular matrix $\mathbf{U}$. Namely, $\mathbf{L}_0\mathbf{X}^t\mathbf{X} = \mathbf{U}$ or $\mathbf{X}^t\mathbf{X} = \mathbf{L}_0^{-1}\mathbf{U} = \mathbf{L}\mathbf{U}$, which is the *LU decomposition* of $\mathbf{X}^t\mathbf{X}$. As $\mathbf{X}^t\mathbf{X}$ is nonnegative definite (n.n.d.), it is better to use its *Cholesky decomposition*. If $\mathbf{X}$ is of full column rank, there exists a unique upper-triangular matrix $\tilde{\mathbf{R}}$ with positive diagonal elements such that $\mathbf{X}^t\mathbf{X} = \tilde{\mathbf{R}}^t\tilde{\mathbf{R}}$. Matrix $\tilde{\mathbf{R}}$ is called the Cholesky factor of $\mathbf{X}^t\mathbf{X}$. The normal equation $\tilde{\mathbf{R}}^t\tilde{\mathbf{R}}\boldsymbol{\beta} = \mathbf{X}^t\mathbf{y}$ can be solved via two equations

$$\tilde{\mathbf{R}}^t\tilde{\mathbf{b}} = \mathbf{X}^t\mathbf{y} \text{ and } \tilde{\mathbf{R}}\boldsymbol{\beta} = \tilde{\mathbf{b}},$$

which can be efficiently solved by back-substitution as $\tilde{\mathbf{R}}$ is upper-triangular. In order to compute other quantities, it is more convenient to consider Cholesky

 © 2012 Wiley Periodicals, Inc.

decomposition of $\mathbf{X}_\mathbf{a}^\mathrm{t}\mathbf{X}_\mathbf{a}$, where $\mathbf{X}_\mathbf{a}$ is the augmented matrix $(\mathbf{X}, \mathbf{y})$ that appends $\mathbf{y}$ as an additional column to $\mathbf{X}$.

QR Decomposition of X

The second method is the QR decomposition of $\mathbf{X}$, i.e.,

$$\mathbf{X} = \mathbf{QR},$$

where $\mathbf{Q}$ is $n \times (p+1)$ with orthonormal columns and $\mathbf{R}$ is $(p+1) \times (p+1)$ upper triangular matrix. If $\mathbf{X}$ is of full column rank, then $\hat{\boldsymbol{\beta}} = \mathbf{R}^{-1}\mathbf{Q}^\mathrm{t}\mathbf{y}$. However, if $\mathbf{X}$ is of rank $m < \min(n, p+1)$, then

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_1 \\ \mathbf{O} \end{bmatrix},$$

where $\mathbf{R}_1$ is $m \times (p+1)$ upper triangular matrix and $\mathbf{O}$ is zero matrix of dimension $(p+1-m) \times (p+1)$. This, together with its corresponding partition of $\mathbf{Q}$, leads to

$$\mathbf{X} = \mathbf{QR} = [\mathbf{Q}_1 \mid \mathbf{Q}_2] \begin{bmatrix} \mathbf{R}_1 \\ \mathbf{O} \end{bmatrix} = \mathbf{Q}_1\mathbf{R}_1, \qquad (46)$$

where $\mathbf{Q}_1$ is an $n \times m$ matrix with orthonormal columns. This form is called a 'skinny' or 'thin' QR, which is more commonly used than the full QR decomposition.

Given the QR decomposition or its thin version of $\mathbf{X}$, the Gram matrix becomes

$$\mathbf{X}^\mathrm{t}\mathbf{X} = \mathbf{R}_1{}^\mathrm{t}\mathbf{Q}_1^\mathrm{t}\mathbf{Q}_1\mathbf{R} = \mathbf{R}_1^\mathrm{t}\mathbf{R}_1,$$

which provides the Cholesky decomposition of $\mathbf{X}^\mathrm{t}\mathbf{X}$.

Three common methods are available for obtaining the QR decomposition of $\mathbf{X}$: the Gram–Schmidt orthogonalization algorithm for $\mathbf{X}$, the Householder transformation, and the Givens transformation. Both Householder and Givens transformations consist of premultiplying $\mathbf{X}$ by orthogonal matrices that transform $\mathbf{X}$ into the upper-triangular matrix $\mathbf{R}$.

SVD Decomposition of X

The singular value decomposition (SVD) of $\mathbf{X}$ has the following form

$$\mathbf{X} = \mathbf{UDV}^\mathrm{t}, \qquad (47)$$

where both $\mathbf{U}_{n\times(p+1)}$ and $\mathbf{V}_{(p+1)\times(p+1)}$ are orthogonal with $\mathbf{U}^\mathrm{t}\mathbf{U} = \mathbf{V}^\mathrm{t}\mathbf{V} = \mathbf{I}_{p+1}$; the columns of $\mathbf{U}$ span the column space of $\mathbf{X}$, $\mathcal{C}(\mathbf{X}) = \mathcal{C}(\mathbf{U})$; the columns of $\mathbf{V}$ span the row space of $\mathbf{X}$, $\mathcal{C}(\mathbf{X}^\mathrm{t}) = \mathcal{C}(\mathbf{V})$; $\mathbf{D}_{(p+1)\times(p+1)} = \mathrm{diag}(d_j)$ is diagonal with entries $d_1 \geq d_j \geq \cdots \geq d_{(p+1)} \geq 0$. The most widely used algorithm for computing SVD, given by Golub and Reinsch,[58] involves a series of Householder transformations and a QR procedure. The algorithm is very stable, but can be prohibitively slow for large n and/or p. Assuming $n \gg p$, the computation time for matrix $\mathbf{X}_{n\times(p+1)}$ is of order $O(np^2)$ floating-point operations (flops).

Given SVD of $\mathbf{X}$ in Eq. (47), it follows that

$$\mathbf{X}^\mathrm{t}\mathbf{X} = \mathbf{VDU}^\mathrm{t}\mathbf{UDV}^\mathrm{t} = \mathbf{VD}^2\mathbf{V}^\mathrm{t},$$

which is the spectral decomposition of $\mathbf{X}^\mathrm{t}\mathbf{X}$. From this result, the principal components of $\mathbf{X}$ can be easily extracted as $\mathbf{XV} = \mathbf{UD}$. If $\mathbf{X}$ is of full column rank, then $\hat{\boldsymbol{\beta}} = \mathbf{VD}^{-1}\mathbf{U}^\mathrm{t}\mathbf{y}$ and the hat matrix $\mathbf{H} = \mathbf{UU}^\mathrm{t}$.

CONCLUSION

In this article, we have gone through some essential concepts and procedures in linear regression and their properties. Linear regression plays an important role in statistics. A thorough understanding of linear model theories and computation is crucial for gaining insight into many of its extensions and new advances. We recommend Kutner et al.[59] for a full account of linear regression from the perspective of application and Seber and Lee[57] for theoretical treatments. A short list of developments that are closely related to linear regression includes generalized linear models, nonlinear least squares, panel data analysis and generalized linear mixed-effects models, mixture regression models, generalized estimating equations and generalized method of moments estimators, time series analysis and dynamic regression models, spatial data analysis, zero-inflated and extreme value regressions, compositional data analysis, multivariate regression and seemingly unrelated regression equations models, structural equation models, quantile regression models, varying coefficient regression models, principal components regression and partial least squares, functional data analysis, local polynomial regression and spline regression, single-index and semiparametric regression models, graphical and social network models, longitudinal data analysis and survival regression models, error-in-variable regression models, and so on. Relatively new advances include artificial neural networks, recursive partitioning, support vector machines, regularization, and ultrahigh dimensional regression models, and so on. Interested readers are referred to Hastie et al.[60] for exposure to these relatively new techniques. In one way or another, they all have some roots that trace back to linear regression.

© 2012 Wiley Periodicals, Inc.

We note that this article is far from being a comprehensive survey of all aspects and approaches involved in linear regression. For example, we have completely omitted the Bayesian approaches to linear regression. One can refer to Box and Tiao,[61] Broemeling,[62] and Congdon[63] for the Bayesian linear model as well as Raftery et al.[64] for the Bayesian model averaging method and Robert and Casella[65] for the Markov Chain Monte Carlo method. In addition, we have not discussed inverse regression analysis for dimension reduction,[66] and a useful reference can be found in Cook.[67] Moreover, we have kept the coverage of analysis of variance short by treating it as a special case of regression. One may refer to Scheffé[3] for a full account. When assessing the effect of treatment on response is of the primary interest, it is critical to distinguish between experimental data and observational data. We refer interested readers to Wu and Hamada[68] for design and analysis of experimental data and Rosenbaum[69] for methods with observational studies. Besides, multiple comparisons,[70] bootstrap method,[71,72] missing data,[73] categorical data,[74,75] gene expression microarray data,[76] financial data,[77] and regression trees[78,79] are among other very important topics related to regression that have also been omitted from this article.

REFERENCES

1. Lehmann EL, Scheffé H. Completeness, similar regions, and unbiased estimation. I. *Sankhyā* 1950, 10:305–340.
2. Lehmann EL, Scheffé H. Completeness, similar regions, and unbiased estimation. II. *Sankhyā* 1955, 15:219–236.
3. Scheffé H. *The Analysis of Variance*. New York: John Wiley & Sons; 1959.
4. Linhart H, Zucchini W. *Model Selection*. New York: Wiley Interscience; 1986.
5. McQuarrie ADR, Tsai CL. *Regression and Time Series Model Selection*. Singapore: World Scientific; 1998.
6. Burnham KP, Anderson DR. *Model Selection and Multi-Model Inference: A Practical Information-Theoretical Approach*. 2nd ed. New York: Springer-Verlag; 2002.
7. Claeskens G, Hjort NL. *Model Selection and Model Averaging*. New York: Cambridge; 2008.
8. Konishi S, Kitagawa G. *Information Criteria and Statistical Modeling*. New York: Springer-Verlag; 2008.
9. Allen DM. The relationship between variable selection and data augmentation and a method for prediction. *Technometrics* 1974, 16:25–127.
10. Craven P, Wahba G. Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the methods of generalized cross-validation. *Numer Math* 1979, 31:377–403.
11. Mallows CL. Some comments on C_p. *Technometrics* 1973, 15:661–675.
12. Akaike H. Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csaki F, eds. *Proceedings of Second International Symposium on Information Theory*. Budapest: Akademiai Kiado; 1973, 267–281.
13. Hurvich CM, Tsai CL. Regression and time series model selection in small samples. *Biometrika* 1989, 76:297–307.
14. Schwarz G. Estimating the dimension of a model. *Ann Stat* 1978, 6:461–464.
15. Kuha J. AIC and BIC: comparisons of assumptions and performance. *Sociol Methods Res* 2004, 33:188–229.
16. Burnham KP, Anderson DR. Multimodel inference: understanding AIC and BIC in model selection. *Sociol Methods Res* 2004, 33:261–304.
17. Yang Y. Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika* 2005, 92:937–950.
18. SAS Institute Inc. *SAS 9.3 Help and Documentation*. Cary, NC: SAS Institute Inc.; 2011.
19. Tibshirani R. Regression shrinkage and selection via the Lasso. *J R Stat Soc Ser B* 1996, 58:267–288.
20. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 1970, 12:55–67.
21. Casella G. Minimax ridge regression estimation. *Ann Stat* 1980, 8:1036–1056.
22. Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression (with discussion). *Ann Stat* 2004, 32:407–499.
23. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc* 2001, 96:1348–1360.
24. Wang H, Li R, Tsai CL. Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* 2007, 94:553–568.
25. Tibshirani R. Regression shrinkage and selection via the Lasso: a retrospective (with comments). *J R Stat Soc Ser B* 2011, 73:273–282.
26. Fan J, Li R. Statistical challenges with high dimensionality: Feature selection in knowledge discovery. In: Sanz-Sole M, Soria J, Varona JL, Verdera J, eds.

 © 2012 Wiley Periodicals, Inc.

Proceedings of the International Congress of Mathematicians, Vol III. Freiburg: European Mathematical Society; 2006, 595–622.

27. Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space (with discussion). *J R Stat Soc Ser B* 2008, 70:849–911.
28. Wang H. Forward regression for ultra-high dimensional variable screening. *J Am Stat Assoc* 2009, 104:1512–1524.
29. Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). *Biometrika* 1965, 52:591–611.
30. Wald A, Wolfowitz J. On a test whether two samples are from the same population. *Ann Math Stat* 1940, 11:147–162.
31. Durbin J, Watson GS. Testing for serial correlation in least squares regression, I. *Biometrika* 1950, 37:409–428.
32. Durbin J, Watson GS. Testing for serial correlation in least squares regression, II. *Biometrika* 1951, 38:159–179.
33. White H. A heteroscedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity. *Econometrica* 1980, 48:817–838.
34. Cook RD, Weisberg S. Diagnostics for heteroscedasticity in regression. *Biometrika* 1983, 70:1–10.
35. Cook RD. Assessment of local influence (with discussion). *J R Stat Soc Ser B* 1986, 48:133–169.
36. Zhu H, Ibrahim JG, Lee S, Zhang H. Perturbation selection and influence measures in local influence analysis. *Ann Stat* 2007, 35:2565–2588.
37. Cook RD. Detection of influential observation in linear regression. *Technometrics* 1977, 15:15–18.
38. Muller KE, Mok MC. The distribution of Cook's *D* statistic. *Commun Stat: Theory Methods* 1997, 26:525–546.
39. Weisberg S. *Applied Linear Regression*. 3rd ed. New York: Wiley Interscience; 2005.
40. Hoaglin D, Velleman P. A critical look at some analyses of Major League Baseball salaries. *Am Stat* 1995, 49:277–284.
41. Box GEP, Tidwell PW. Transformations of independent variables. *Technometrics* 1962, 4:531–550.
42. Box GEP, Cox DR. An analysis of transformations (with discussion). *J R Stat Soc Ser B* 1964, 26:211–246.
43. Carroll RJ, Ruppert D. *Transformations and Weighting in Regression*. New York: Chapman and Hall; 1988.
44. Atkinson AC. *Plots, Transformations and Regression*. Oxford: Oxford University Press; 1985.
45. Bickel P, Doksum K. An analysis of transformations revisited. *J Am Stat Assoc* 1982, 77:296–311.
46. Carroll RJ, Ruppert D. On prediction and the power transformation family. *Biometrika* 1981, 68:609–615.
47. Box GEP, Cox DR. An analysis of transformations revisited, rebutted. *J Am Stat Assoc* 1982, 77:209–210.
48. Stone CJ. Additive regression and other nonparametric models. *Ann Stat* 1985, 13:689–705.
49. Hastie TJ, Tibshirani RJ. *Generalized Additive Models*. New York: Chapman and Hall; 1990.
50. Breiman L, Friedman JH. Estimating transformations for multiple regression and correlation. *J Am Stat Assoc* 1985, 80:580–598.
51. Tibshirani R. Estimating transformations for regression via additivity and variance stabilization. *J Am Stat Assoc* 1988, 83:394–405.
52. Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics* 1982, 38:963–974.
53. Demidenko E. *Mixed Models: Theory and Applications*. New York: Wiley Interscience; 2004.
54. Huber PJ. *Robust Statistics*. New York: John Wiley & Sons; 1981.
55. Hampel FR, Ronchetti EM, Rousseeuw PJ, Stahel WA. *Robust Statistics: The Approach Based on Influence Functions*. New York: Wiley Interscience; 1986.
56. Gentle JE. *Computational Statistics*. New York, NY: Springer-Verlag; 2009.
57. Seber GAF, Lee AJ. *Linear Regression Analysis*. 2nd ed. New York: Wiley Interscience; 2003.
58. Golub GH, Reinsch C. Singular value decomposition and least squares solutions. In: Wilkinson JH, Reinsch C, eds. *Handbook for Automatic Computation, Vol. 1 (Linear Algebra)*. New York: Springer-Verlag; 1971, 134–151.
59. Kutner M, Nachtsheim C, Neter J, Li W. *Applied Linear Statistical Models*. 5th ed. Boston, MA: McGraw-Hill/Irwin; 2004.
60. Hastie T, Tibshirani R, Friedman JH. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer-Verlag; 2009.
61. Box GP, Tiao GC. *Bayesian Inference in Statistical Analysis*. MA: Addison-Wesley; 1973.
62. Broemeling LD. *Bayesian Analysis of Linear Models*. New York: Marcel Dekker; 1985.
63. Congdon P. *Bayesian Statistical Modelling*. New York: Wiley Interscience; 2001.
64. Raftery AE, Madigan D, Hoeting JA. Bayesian model averaging for linear regression models. *J Am Stat Assoc* 1997, 92:179–191.
65. Robert CP, Casella G. *Monte Carlo Statistical Methods*. 2nd ed. New York: Springer-Verlag; 2004.
66. Li KC. Sliced inverse regression for dimension reduction (with discussion). *J Am Stat Assoc* 1991, 86:316–342.
67. Cook RD. *Regression Graphics: Ideas for Studying Regressions through Graphs*. New York: Wiley Interscience; 1998.

 © 2012 Wiley Periodicals, Inc.

68. Wu CFJ, Hamada M. *Experiments: Planning, Analysis, and Parameter Design Optimization*. New York: Wiley Interscience; 2000.
69. Rosenbaum PR. *Observational Studies*. 2nd ed. New York: Springer-Verlag; 2010.
70. Hochberg Y, Tamhane AC. *Multiple Comparison Procedures*. New York: Wiley Interscience; 1987.
71. Efron B, Tibshirani R. *An Introduction to the Bootstrap*. London: Chapman and Hall; 1993.
72. Davison AC, Hinkley DV. *Bootstrap Methods and Their Application*. New York: Cambridge University Press; 1997.
73. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. 2nd ed. New York: Wiley Interscience; 2002.
74. Agresti A. *Categorical Data Analysis*. 2nd ed. New York: Wiley Interscience; 2002.
75. Simonoff JS. *Analyzing Categorical Data*. New York: Springer-Verlag; 2003.
76. McLachlan GJ, Do KA, Ambroise C. *Analyzing Microarray Gene Expression Data*. New York: Wiley Interscience; 2004.
77. Tsay RS. *Analysis of Financial Time Series*. 2nd ed. New York: Wiley Interscience; 2005.
78. Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and Regression Trees*. Boca Raton, FL: Chapman & Hall/CRC; 1984.
79. Su X, Tsai CL, Yan X. Treed variance. *J Comput Graph Stat* 2006, 15:356–371.

FURTHER READING

Anderson TW. *An Introduction to Multivariate Statistical Analysis*. 2nd ed. New York: John Wiley & Sons; 1984.

Björck A. *Numerical Methods for Least Squares Problems*. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM); 1996.

Brockwell PJ, Davis RA. *Time Series: Theory and Methods*. 2nd ed. New York: Springer-Verlag; 1991.

Chatterjee S, Hadi AS. *Sensitivity Analysis in Linear Regression*. New York: Wiley Interscience; 1988.

Cox DR, Hinkley DV. *Theoretical Statistics*. Boca Raton, FL: Chapman & Hall/CRC Press; 1974.

Cox DR. Regression models and life tables (with discussion). *J R Stat Soc Ser B* 1972, 34:187–220.

Cressie NAC. *Statistics for Spatial Data*. New York: Wiley Interscience; 1993.

Diggle PJ, Heagerty P, Liang K-Y, Zeger SL. *Analysis of Longitudinal Data*. 2nd ed. Oxford: Oxford University Press; 2002.

Eubank RL. *Spline Smoothing and Nonparametric Regression*. New York: Marcel Dekker; 1988.

Fan J, Gijbels I. *Local Polynomial Modelling and Its Applications*. London: Chapman and Hall; 1996.

Fuller WA. *Measurement Error Models*. New York: John Wiley & Sons; 1987.

Hamilton JD. *Time Series Analysis*. Princeton, NJ: Princeton University Press; 1994.

Hsu JC. *Multiple Comparisons: Theory and Methods*. London: Chapman and Hall; 1996.

Klein JP, Moeschberger ML. *Survival Analysis: Techniques for Censored and Truncated Data*. 2nd ed. New York: Springer-Verlag; 2003.

Koenker R. *Quantile Regression*. New York: Cambridge University Press; 2005.

Lawless JF. *Statistical Models and Methods for Lifetime Data*. 2nd ed. New York: John Wiley & Sons; 2003.

McCullagh P, Nelder JA. *Generalized Linear Models*. London: Chapman and Hall; 1989.

McLachlan G, Peel D. *Finite Mixture Models*. New York: Wiley Interscience; 2000.

Ramsey JO, Silverman BW. *Functional Data Analysis*. 2nd ed. New York: Springer-Verlag; 2005.

Rao CR. *Linear Statistical Inference and its Applications*. 2nd ed. New York: John Wiley & Sons; 1973.

Rossi PE, Allenby GM, McCulloch R. *Bayesian Statistics and Marketing*. New York: Wiley Interscience; 2005.

Schott JR. *Matrix Analysis for Statistics*. 2nd ed. New York: John Wiley & Sons; 2005.

Seber GAF, Wild CJ. *Nonlinear Regression*. New York: Wiley Interscience; 1989.

Simonoff JS. *Smoothing Methods in Statistics*. New York: Springer-Verlag; 1996.

Yan X, Su XG. *Linear Regression Analysis: Theory and Computing*. Hackensack, NJ: World Scientific; 2009.

 © 2012 Wiley Periodicals, Inc.