Chapter 2.

Linear Models

This chapter of introducing the linear regression model emphasizes geometric interpretation and may be abstract but quite condensed and informative. Consider linear regression model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \qquad i = 1, ..., n \tag{2.1}$$

where $y_i, x_i = (x_{i1}, ..., x_{ip}))$ are the i-th observation of the response and covariates. Responses are some times called dependent variables or outputs; and the covariates called independent variables or inputs or regressors. Our aim is to through analysis of the data to obtain the parameter estimation and making prediction of any responses on given covariates. The geometric interpretation depends on a fact that the zero-correlation of two variables from multivariate normal random variable implies their independence. Suppose $\mathbf{z} = (z_1, ..., z_n)^T$ are z_i are iid standard normal random variables. Let $\mathbf{z}_1 = \mathbf{A}\mathbf{z}$ and $\mathbf{z}_2 = \mathbf{B}\mathbf{z}$ with $\mathbf{A}$ and $\mathbf{B}$ are two nonrandom matrices. Then

$$\text{cov}(\mathbf{z}_1, \mathbf{z}_2) = \mathbf{A}\mathbf{B}^T = 0$$

implies the independence between $\mathbf{z}_1$ and $\mathbf{z}_2$. Then we also call $\mathbf{z}_1$ and $\mathbf{z}_2$ orthogonal.

2.1. The least squares approach and geometric properties.

With slight abuse of notation, in this chapter, we use

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & ... & x_{1p} \\ 1 & x_{21} & x_{22} & ... & x_{2p} \\ \vdots & \vdots & \vdots & ... & \vdots \\ 1 & x_{n1} & x_{n2} & ... & x_{np} \end{pmatrix} = \left(\mathbf{1} \vdots \mathbf{x}_1 \vdots \mathbf{x}_2 \vdots \cdots \vdots \mathbf{x}_p \right).$$

Here a column of ones, $\mathbf{1}$, is added, which corresponds to the intercept β_0. Then $\mathbf{X}$ is a n by $p+1$ matrix. Recall that

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_n \end{pmatrix}, \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}, \quad \mathbf{x}_j = \begin{pmatrix} 1 \\ \vdots \\ x_{nj} \end{pmatrix},$$

The least squares criterion is try to minimize the sum of squares:

$$\sum_{i=1}^{n} (y_i - \beta_0 + \beta_1 x_{i1} + \cdots + x_{ip})^2.$$

Using matrix algebra, the above sum of squares is

$$\|\mathbf{y} - \mathbf{X}\beta\|^2 = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta).$$

By some linear algebra calcuation, the least squares estimator of β is then

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

Then

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$$

is called the fitted values. They can be viewed as the predicted values of the reponses based on the linear model. And

$$\mathbf{y} - \hat{\mathbf{y}}$$

are called residuals, which is denoted as $\hat{\epsilon}$. The sum of squares of these residuals

$$\sum_{i=1}^{n} \hat{\epsilon}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \|\mathbf{y} - \hat{\mathbf{y}}\|^2.$$

A key observation is that the residual $\hat{\epsilon}$ is orthogonal to all columns of $\mathbf{X}$, i.e, all $\mathbf{1}, \mathbf{x}_1, ..., \mathbf{x}_p$. This can be seen by

$$\begin{aligned} & \mathbf{X}^T \hat{\epsilon} = \mathbf{X}^T \mathbf{y} - \mathbf{X}^T \mathbf{X} \hat{\beta} \\ = \; & \mathbf{X}^T \mathbf{y} - \mathbf{X}^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} = 0. \end{aligned}$$

In other words the residual vector $\hat{\epsilon}$ is orthogonal to the hyperplane formed by vectors $\mathbf{1}, \mathbf{x}_1, ..., \mathbf{x}_p$ in n dimensional real space. This fact can also be used to prove that $\hat{\beta}$ is the least squares estimator, since

$$\begin{aligned} & \|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2 \\ = \; & \|\mathbf{y} - \mathbf{X}\hat{\beta} - \mathbf{X}(\mathbf{b} - \hat{\beta})\|^2 \\ = \; & \|\mathbf{y} - \mathbf{X}\hat{\beta}\|^2 + \|\mathbf{X}(\mathbf{b} - \hat{\beta})\|^2 \qquad \text{by orthogonality} \\ \geq \; & \|\mathbf{y} - \mathbf{X}\hat{\beta}\|^2 \end{aligned}$$

Notice that the fitted value $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$, which, also as a vector in n dimensional real space, is a linear combination of the vectors $\mathbf{1}, \mathbf{x}_1, ..., \mathbf{x}_p$, with the $p + 1$ linear combination coefficients being the components of $\hat{\beta}$. As a result, the fitted values are orthogonal to the residuals, i.e., $\hat{\mathbf{y}}$ is orthogonal to $\mathbf{y} - \hat{\mathbf{y}}$ or

$$\hat{\mathbf{y}}^T (\mathbf{y} - \hat{\mathbf{y}}) = 0.$$

This implies

$$\|\mathbf{y}\|^2 = \|\mathbf{y}\|^2 + \|\mathbf{y} - \hat{\mathbf{y}}\|^2.$$

The projection matrix. Let $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. This n by n matrix is called projection matrix or hat matrix. It has the property that, for any vector, $\mathbf{b}$ in n dimensional real space $\mathbf{Hb}$ projects $\mathbf{b}$ onto the linear space formed by the columns of $\mathbf{X}$. It means that $\mathbf{Hb}$ is in this linear space formed by the columns of $\mathbf{X}$. And $\mathbf{b} - \mathbf{Hb}$ is orthogonal to this space.

In general, a projection matrix $\mathbf{H}$ is symmetric and idempotent; i.e., $\mathbf{H}^2 = \mathbf{H}$. In other words, it is a matrix with eigenvalues being either 1 or 0. All eigenvectors associated with eigenvalue 1 form a space, say $\mathcal{L}_1$; and those with eigenvalue 0 form the orthogonal space, $\mathcal{L}_0$, of $\mathcal{L}_1$. Then $\mathbf{H}$ is the projection onto space $\mathcal{L}_1$ and $\mathbf{I} - \mathbf{H}$ is the projection onto $\mathcal{L}_0$, where $\mathbf{I}$ is the n by n identity matrix.

Matrix decomposition Two facts in linear algebra shall be very useful: Suppose, for convenience $n \geq p$, any matrix n by$p$ matrix A can always be decomposed into

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^T$$

where $\mathbf{U}$ is $n \times p$ orthogonal matrix, $\mathbf{D}$ is a $p \times p$ diagonal matrix and $\mathbf{V}$ is $p \times p$ orthogonal matrix. In particular

$$\mathbf{X} = \mathbf{U}\mathbf{R},$$

where $\mathbf{R} = \mathbf{D}\mathbf{V}$.

If $\mathbf{A}$ and $\mathbf{B}^T$ are two matrices of same dimension, then

$$trace(\mathbf{AB}) = trace(\mathbf{BA}).$$

2.2. The statistical properties of the least squares estimates.

The linear regression model general assumes the error ϵ_i has zero conditional mean and constant conditional variance σ^2, and the covariates x_i are non-random; and there are independence across the observations $(x_i, \epsilon_i), i = 1, ..., n$. Often a more restrictive condition is assumed: the error follow normal distribution, i.e, $N(0, \sigma^2)$. Let $c_{00}, c_{11}, ..., c_{pp}$ be the diagonal elements of $(\mathbf{X}^T\mathbf{X})^{-1}$. Under the above conditions,

$$\begin{aligned}
&\hat{\beta} \sim N(\beta, \sigma^2(\mathbf{X}\mathbf{X}^T)^{-1});\\
&\text{RSS} = \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \sim \sigma^2 \chi^2_{n-p-1}\\
&\hat{\beta} \text{ and RSS are independent}\\
&s^2 = \text{RSS}/(n-p-1) \text{ unbiased estimate of } \sigma^2\\
&\frac{\hat{\beta}_j - \beta_j}{s\sqrt{c_{jj}}} \sim t_{n-p-1}\\
&\frac{(\hat{\beta} - \beta)^T(\mathbf{X}^T\mathbf{X})(\hat{\beta} - \beta)/p}{s^2} \sim F_{p+1,n-p-1}
\end{aligned}$$

To understand the above result, a key observation is that

$$\begin{aligned}
\text{cov}(\hat{\beta}, \hat{\epsilon}) &= \text{cov}((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon, (\mathbf{I} - \mathbf{H})\epsilon))\\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \text{var}(\epsilon)(\mathbf{I} - \mathbf{H}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{I} - \mathbf{H})\\
&= 0
\end{aligned}$$

because $\mathbf{H}$ is idempotent. These results can be used to construct confidence intervals as well as test of hypothesis. For example,

$$\hat{\beta}_j \pm t_{n-p-1}(\alpha/2)s\sqrt{c_{jj}}$$

is a confidence interval for β_j at confidence level $1 - \alpha$. Here $t_{n-p-1}(\alpha/2)$ is the $1 - \alpha/2$ percentile of the t-distribution with degree of freedom $n - p - 1$.

For a given value of input $\mathbf{x}$, its mean response is $\beta^T\mathbf{x}$. The confidence interval this mean response is

$$\hat{\beta}^T\mathbf{x} \pm t_{n-p-1}(\alpha/2)s\sqrt{\mathbf{x}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}}$$

The confidence interval for β_j is a special case of the above formula by taking $\mathbf{x}$ as a vector that all zero except the $(j + 1)$ entry corresponding β_j. (Because of β_0, β_j is at the $j + 1$th position of $\hat{\beta}$.

To predict the actual response y, rather than its mean, we would use the same point estimator $\hat{\beta}^T\mathbf{x}$, but the accuracy is much decreased as more uncertainty in the randomness of the actual response from the error is involved. The confidence interval, often called prediction interval, for y is

$$\hat{\beta}^T\mathbf{x} \pm t_{n-p-1}(\alpha/2)s\sqrt{1 + \mathbf{x}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}}.$$

2.3. The variance decomposition and analysis of variance (ANOVA).

Recall that

$$\|\mathbf{y}\|^2 = \|\hat{\mathbf{y}}\|^2 + \|\mathbf{y} - \hat{\mathbf{y}}\|^2.$$

The common variance decomposition takes a similar form, but leaving out sample mean, :

$$\|\mathbf{y} - \bar{y}\|^2 = \|\hat{\mathbf{y}} - \bar{y}\|^2 + \|\mathbf{y} - \hat{\mathbf{y}}\|^2;$$

which is often written as

$$\text{SS}_{total} = \text{SS}_{reg} + \text{SS}_{error}.$$

The quantity on the left of equation, the total sum of squares, measures the total variation in response. The middle quantity, the sum of squares due to regression or, more precisely, due to the

inputs, measures variation in response explained by that of the inputs. The last quantity, the sum of squares due to error, measures the size of randomness due to error or noise. The ANOVA table looks like

Source of Variation	SumOfSquares	Degree of Freedom	Mean Squared	F-statistic
Regression	SS_{reg}	p	MS_{reg}	$\text{MS}_{reg}/\text{MS}_{error}$
Error	SS_{error}	$n-p-1$	MS_{error}	
Total	SS_{total}	$n-1$		

where $\text{MS}_{reg} = \text{SS}_{reg}/p$ and $\text{MS}_{error} = \text{SS}_{error}/(n-p-1)$. And the F-statistic follows $F_{p,n-p-1}$ distribution under the hypothesis that $\beta_1 = \beta_2 = ...\beta_p = 0$, i.e., all inputs are unrelated with the output. The p-value is the probability for the distribution $F_{p+1,n-p-1}$ taking value greater than the value of the F-statistic.

The above ANOVA is a special case of a general variance decomposition. Let $\mathcal{L}$ be the linear space spanned by $\mathbf{1}, \mathbf{x}_1, ..., \mathbf{x}_p$, all columns of $\mathbf{X}$. Note that our linear model assumption:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

can be written as $E(\mathbf{y}) = \mathbf{X}\beta$, or

$$E(\mathbf{y}) \in \mathcal{L}.$$

The fitted values $\hat{\mathbf{y}}$ is projection of $\mathbf{y}$ onto $\mathcal{L}$, and $s^2 = \|\mathbf{y}-\hat{\mathbf{y}}\|^2/(n-p-1)$ is the unbiased estimator of σ^2.

Now let's further assume that

$$E(\mathbf{y}) \in \mathcal{L}_0$$

where $\mathcal{L}_0$ is some linear subspace of $\mathcal{L}$ of dimension $r < p+1$. Let $\hat{\mathbf{y}}_0$ be the project of $\mathbf{y}$ on to $\mathcal{L}_0$. Then, Pythagorean theorem implies

$$\|\mathbf{y} - \hat{\mathbf{y}}_0\|^2 = \|\mathbf{y} - \hat{\mathbf{y}}\|^2 + \|\hat{\mathbf{y}} - \hat{\mathbf{y}}_0\|^2$$

Then, by the same token, $s_0^2 = \|\mathbf{y}-\hat{\mathbf{y}}_0\|^2/(n-r)$ is the unbiased estimator of σ^2 under the hypothesis $E(\mathbf{y}) \in \mathcal{L}_0$.

It can be shown that

$$F = \frac{\|\hat{\mathbf{y}} - \hat{\mathbf{y}}_0\|^2/(p+1-r)}{\|\mathbf{y} - \hat{\mathbf{y}}\|^2/(n-p-1)} \sim F_{p+1-r,n-p-r}$$

This F-statistic is used to test the hypothesis that $H_0 : E(\mathbf{y}) \in \mathcal{L}_0$, against the alternative H_a : otherwise. The commonly considered hypothesis, as dealt with in the ANOVA table, $H_0 : \beta_1 = \cdots = \beta_p = 0$ can be formulated as $H_0 : E(\mathbf{y}) \in \mathcal{L}(\mathbf{1})$, where $\mathcal{L}(\mathbf{1})$ represent the linear space of a single vector $\mathbf{1}$.

In variable selection problems, we may be concerned with a subset of the p variables are irrelevant with the response. Let the subset be denoted as $A = \{i_1, ..., i_r\}$, where $r \leq p$. Then, the null hypothesis is

$$H_0 : \beta_{i_1} = \beta_{i_2} = \cdots = \beta_{i_r} = 0,$$

which again is equivalent to

$$H_0 : E(\mathbf{y}) \in \mathcal{L}(A^c),$$

where $\mathcal{L}(A^c)$ is the linear space in R^n spanned by $\mathbf{x}_j$ for $j \notin A$, which is $p+1-r$ dimension.

2.4. The optimality of the least squares estimation and the Gauss-Markov Theorem.

The least squares has many superior properties, such as easy computation and consistency and efficiency. For estimates of β, suppose we concentrate only on all linear unbiased estimates: $\sum_{i=1}^n \mathbf{a}_i y_i$, with $\mathbf{a}_i$ being nonrandom. Then,

Theorem (Gauss-Markov Theorem). Among all linear unbiased estimates, the least squares estimate has the smallest variance, thus smallest mean squared error.

Proof. Let $\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_n)$. Then, unbiased estimate of β is $\mathbf{Ay}$ with mean $\mathbf{AX}\beta = \beta$ by the unbiasedness, and variance matrix $\mathbf{AA}^T$. Write $\mathbf{A} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T + \mathbf{D}$. Then, $\mathbf{DX} = 0$. As a result,

$$\mathbf{AA}^T = (\mathbf{X}^T\mathbf{X})^{-1} + \mathbf{DD}^T \geq \Big(\mathbf{X}^T\mathbf{X})^{-1}.$$

Here the inequality is for symmetric matrices, i.e., $\mathbf{A} \geq \mathbf{B}$ is defined as $\mathbf{A} - \mathbf{B}$ is nonnegative definite.

2.5. Regression diagnostics.

There are several issues on regression diagnostics, which include but not limited to:

a). Error distribution (normality check). Non-normality may cause the normality-based inference such as t-test and F-test being inaccurate, if the sample size is not large. For the distribution of the error terms, We can use graphics, such histogram, boxplot and qqnorm to visualize the the distribution of the residuals. However, the residuals $\hat{\epsilon}_i$ does not follow the distribution $N(0, \sigma^2)$, even if all model assumptions are correct! Recall that

$$\hat{\epsilon} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{H})\epsilon \sim N(0, \sigma^2(\mathbf{I} - \mathbf{H})).$$

So, $\hat{\epsilon}_i \sim N(0, (1 - h_{ii})\sigma^2)$. An (internally) studentized residual is

$$e_i = \frac{\hat{\epsilon}_i}{s\sqrt{1 - h_{ii}}}.$$

A more detailed check may use the studentized residuals. (A more appropirate one is the (externally) studentized residual which uses an s from the least squares fitting by deleting the i-th observation.)

b). Homoscedasticity versus heteroscedasticity. Heteroscedasticity can cause the estimate being not the optimal one, which may be fixed by weighted least squares estimation. Use scatter plot of residuals against the fitted values to check the heteroscedasticity (the variance of the errors are not equal).

c). Error dependence. IF the errors are dependent, the inference will also be incorrect. Use autocorrelation of the residuals (ACF) to check the independence assumption of the errors. One can also use Durbin-Watson test, which tests whether the first few autocorrelations are 0.

d). Leverage an Cook's D. Recall the hat matrix

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T.$$

Let $h_{ij} = \mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_j$ be the (i, j) elements of $\mathbf{H}$. The leverage of the i-th observation is just the i-th diagonal element of $\mathbf{H}$, denoted as h_{ii}. A high leverage implies that observation is quite influential. Note that the average of h_{ii} is $(p + 1)/n$ (Exercise 2.1). So, if h_{ii} is greater than $2(p + 1)/n$, twice of the average, is generally considered large.

More precisely, the Cook'D is often used measure how important an observation is. Cook's D is defined

$$D_i = \frac{\sum_{k=1}^n (\hat{y}_k - \hat{y}_k^{(-i)})^2}{(p + 1)s^2}$$

where $\hat{y}_k$ is the k-th fitted value; and $\hat{y}_k^{(-i)}$ is the k-th fitted value by deleting the i-th observation. If D_i is large, it implies once i-th observation is not available, the prediction would be much different, thus reflecting the importance of this observation. In general, the observations with large D_i, such as larger than a quarter of the sample size, may be considered influential.

e). Multicollinearity. The multicollinarity can cause the parameter estimation to be very unstable. Suppose two inputs are strongly correlated, their separate effect on regression is difficult to identify from the regression. When data change slightly, the two regression coefficients can differ greatly,

though their joint effect may stay little changed. It is common to use variance inflation factor (VIF) to measure one input's correlation with the others. The largest value of VIF, ∞, means this input is perfectly linearly related with the other inputs. The smallest value of VIF, 1, means this input is uncorrelated with the other inputs. In general, variable selection methods may be used to reduce the number of highly correlated variables.

2.6. Variable selection.

Variable selection, or more generally, model selection, is an important tool in minimizing prediction error. There are substantial research development regarding methods of model selection. The aim is to minimize generalization error or prediction error. The naive approach is to exhaust all models. However, with the curse of dimensionality, this is quickly prohibitive when the number of variables increase. There are more sophisticated methods such as cross validation or regularization methods, such as LASSO (Tibshirani 1996). Here we introduce more basic and simple methods.

More inputs do not imply better prediction, particularly if the inputs in the model are irrelevant with the response. Moreover, more inputs also imply more danger of overfit, resulting in small training error but large test error.

a). Adjusted R-squared.

The R-squared is the $\mathrm{SS}_{Reg}/\mathrm{SS}_{total}$, which is the percentage of the total variation in response due to the inputs. The R-squared is commonly used as a measurement of how good the linear fit is. However, a model with larger R-squared is not necessarily better than another model with smaller R-squared. If model A has all the inputs of model B, then model A's R-squared will always be greater than or as large as that of model B. If model A's additional inputs are entirely uncorrelated with the response, model A contain more noise than model B. As a result, the prediction based on model A would inevitable be poorer or no better.

Recall that the R-squared is defined as:

$$R^2 = \frac{\mathrm{SS}_{reg}}{\mathrm{SS}_{total}} = 1 - \frac{\mathrm{SS}_{error}}{\mathrm{SS}_{total}}$$

where $\mathrm{SS}_{error} = \sum_{i=1}^{n} \hat{\epsilon}_i^2$ is often called residual sum of squares (RSS).

The adjusted R-squared, taking into account of the degrees of freedom, is defined as

$$\begin{aligned} \text{adjusted } R^2 &= 1 - \frac{\mathrm{MS}_{error}}{\mathrm{MS}_{total}} \\ &= 1 - \frac{\mathrm{SS}_{error}/(n-p-1)}{\mathrm{SS}_{total}/(n-1)} \\ &= 1 - \frac{s^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2/(n-1)} \end{aligned}$$

With more inputs, the R^2 always increase, but the adjusted R^2 could decrease since more inputs is penalized by the smaller degree of freedom of the residuals. The adjusted R-squared is preferred over the R-squared in evaluating models.

b). Mallows' C_p.

Recall that our linear model (2.1) has p covariates, and $s^2 = \mathrm{SS}_{error}/(n-p-1)$ is the unbiased estimator of σ^2. Assume now more covariates are available. Suppose we use only p of the K covariates with $K \geq p$. The statistic of Mallow's C_p is defined as

$$\frac{\mathrm{SS}_{error}(p)}{s_K^2} - 2(p+1) - n.$$

where SS_{error} is the residual sum of squares for the linear model with p inputs and s_K^2 is the unbiased estimator of σ^2 based on K inputs. The smaller Mallows' C_p is, the better the model is.

The following AIC is more often used, despite that Mallws' C_p and AIC usually give the same best model.

c). AIC.

AIC stands for Akaike information criterion, which is defined as

$$\text{AIC} = \log(s^2) + 2(1+p)/n,$$

for a linear model with p inputs, where $s^2 = \text{SS}_{error}/(n-p-1)$. AIC aims at maximizing the predictive likelihood. The model with the smallest AIC is preferred.

The AIC criterion is try to maximize the expected predictive likelihood. In general, it can be roughly derived in the following. Let θ be a parameter of d dimension. $\hat{\theta}$ is the maximum likelihood estimator of θ based on observations $y_1, ..., y_n$. Let θ_0 be the true (unknown) value of θ, and $\mathcal{I}(\theta_0)$ be the Fisher information. Then, the (expected) predictive log-likelihood is

$$\begin{aligned} E(\log f(Y|\theta))|_{\theta=\hat{\theta}} &\approx E(\log f(Y|\theta))|_{\theta=\theta_0} - \frac{1}{2}(\hat{\theta}-\theta_0)^T\mathcal{I}(\theta_0)(\hat{\theta}-\theta_0) \\ &\approx \frac{1}{n}\sum_{i=1}^{n}\log f(y_i|\hat{\theta}) - \frac{1}{2}(\hat{\theta}-\theta_0)^T\mathcal{I}(\theta_0)(\hat{\theta}-\theta_0) - \frac{1}{2}(\hat{\theta}-\theta_0)^T\mathcal{I}(\theta_0)(\hat{\theta}-\theta_0) \\ &\approx \frac{1}{n}(\text{maximum log likelihood}) - (\hat{\theta}-\theta_0)^T\mathcal{I}(\theta_0)(\hat{\theta}-\theta_0) \\ &\approx \frac{1}{n}(\text{maximum log likelihood}) - d \end{aligned}$$

The approximations are due to the Taylor expansion. Then, maximizing the above predictive likelihood is the same as minimize

$$-2(\text{maximum log likelihood}) + 2d$$

where, the first term is called deviance. In the case of linear regression with normal errors, the deviance is the same as $\log(s^2)$.

d). BIC.

BIC stands for Schwarz's Bayesian information criterion, which is defined as

$$\text{BIC} = \log(s^2) + (1+p)\log(n)/n,$$

for a linear model with p inputs. Again, the model with the smallest BIC is perferred. The derivation of BIC results from Bayesian statistics and has Bayesian interpretation. It is seen that BIC is formally similar to AIC. The BIC penalizes more heavily the models with more number of inputs.

2.6. Examples

Example 2.1 ⋆⋆ In Chapter 1, there is a plot about the closing price of Shanghai index from year 2001 to 2016. We observe that there are two major bull-market periods during the last 16 years, which happened in 2007 and 2015. It's often said that the 2007 bull market is a feast for blue chips while the 2015 bull market favors middle and small size stocks more. Is that right?

Here we do a univariate linear regression of stock returns on capitalization, which describes the scale of a stock. We consider two half-year periods in 2007 (April 17, 2015 to Oct 16, 2015) and 2015 (December 13, 2014 to June 12, 2015) separately. The response is the half-year return for all listed stocks, and the covariate is their capitalization at the beginning of each period (the unit is 100 million RMB). The result is shown in Table 1.1.

Observe that the parameters of cap are very small. That is because the numbers of capitalization are usually too large for stock returns. To get a more clear result, we do a log transformation on the covariate (named as log.cap). The result is shown below.

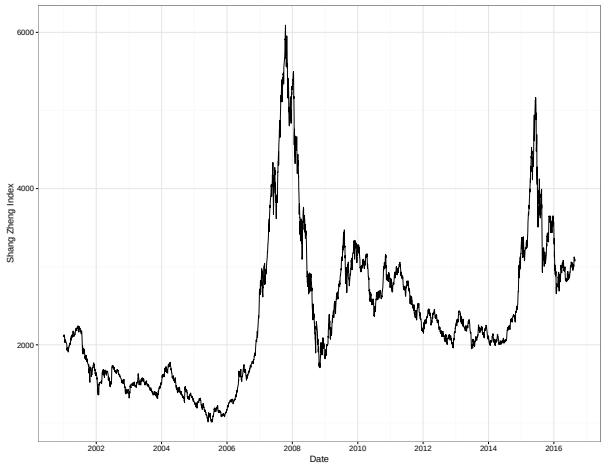


Figure 1: The Shanghai Index from 2000-2016

	Dependent variable:	
	ret	
	Period 1	Period 2
cap	0.0001**	−0.0001***
	(0.00002)	(0.00002)
Constant	0.410***	1.242***
	(0.016)	(0.016)
Observations	1,402	2,574
R^2	0.004	0.014
Adjusted R^2	0.003	0.014
Residual Std. Error	0.602 (df = 1400)	0.779 (df = 2572)
F Statistic	5.715** (df = 1; 1400)	36.596*** (df = 1; 2572)
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 1: Regression Results of ret ∼ cap

	Dependent variable:	
	ret	
	Period 1	Period 2
log.cap	0.112***	−0.230***
	(0.015)	(0.016)
Constant	0.007	2.200***
	(0.057)	(0.070)
Observations	1,402	2,574
R^2	0.039	0.075
Adjusted R^2	0.038	0.074
Residual Std. Error	0.592 (df = 1400)	0.755 (df = 2572)
F Statistic	56.284*** (df = 1; 1400)	207.738*** (df = 1; 2572)
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 2: Regression Results of ret ∼ log.cap

By the results of regression, we could see in Period 1 the log.cap is positive correlated with the response return, while in Period 2 they are negative correlated. Note that the p-values of the coefficients of log.cap are less than 0.05, therefore, the covariate log.cap is significantly at level 0.05. The C.Is of coefficient of log.cap at level 0.05 are $(0.0828, 0.1414)$ and $(-0.2616, -0.1989)$ respectively.

We can also see this point visually from the following plot.

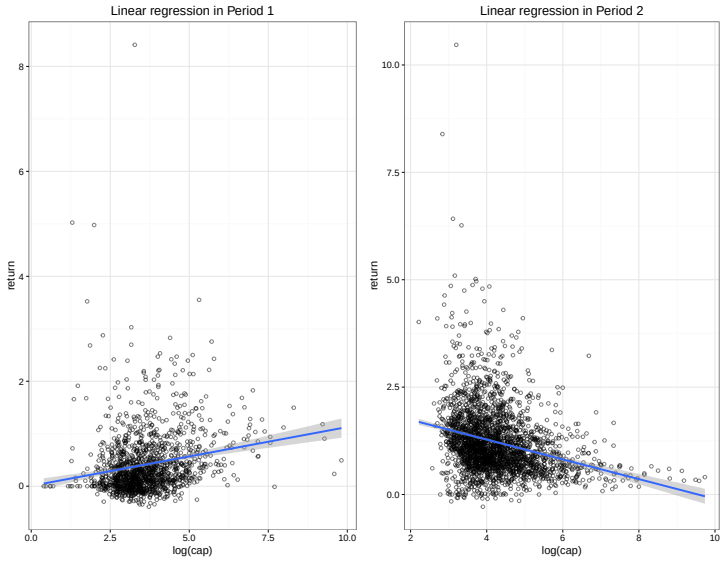


Figure 2: Scatter Plots for Period 1 and Period 2

Also, concerning the leverages, Cook's D and standardized residuals, we plot the following figures.

It can be seen from the q-q plot that the standardized residual is not normal. The fitted model have low leverage and high residuals and all the observations are not influential. Moreover, the plot of Scaled-locations shows that the observation nearly meet the assumption of homoscedasticity.

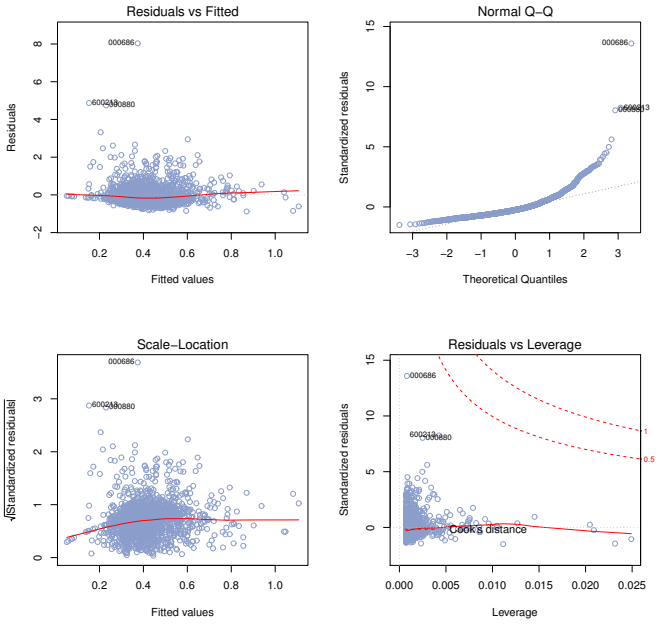


Figure 3: Plots of fitted model in Period 1

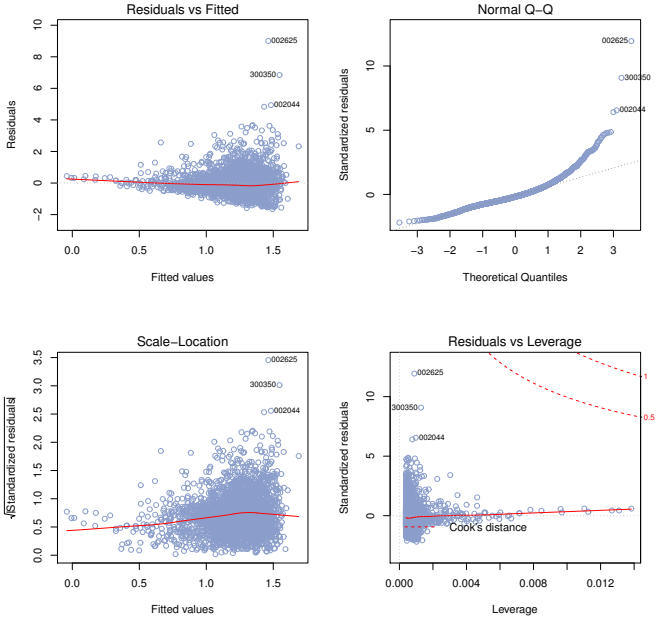


Figure 4: Plots of fitted model in Period 2

Example 2.2 ⋆⋆ Now let's consider a multivariate example. In A-stock market, there are three companies **SH600489**, **SZ002716** and **SH600459** who produce gold (**Au**), silver (**Ag**) and platinum (**Pt**) as their major business correspondingly. So their stock price must be highly correlated to the spot price of Au, Ag and Pt. In this example, we consider three stocks separately. We choose daily returns as response, and Au, Ag and Pt's daily returns as covariates. To get more information, daily returns of U.S. Dollar Index (named as **US**) and Shanghai Index (named as **SH**) are also involved in the model. It is a common sense that U.S. Dollar Index is negatively related to the price of precious metal and ShangHai Index has a great influence on all listed stocks in China A-stock market.

Then, we can build multivariate linear models such as

$$\textbf{SH600489 / SZ002716 / sh600459} = \beta_0 + \beta_1 \textbf{Au} + \beta_2 \textbf{Ag} + \beta_3 \textbf{Pt} + \beta_4 \textbf{US} + \beta_5 \textbf{SH}$$

The result is shown in the following table.

Table 3: Regression Results of Example 2.2

	Dependent variable:		
	sh600489	sz002716	sh600459
au	0.319***	0.443*	0.067
	(0.069)	(0.233)	(0.065)
ag	−0.064	−0.048	−0.042
	(0.040)	(0.146)	(0.037)
pt	0.085*	0.298*	0.075*
	(0.045)	(0.159)	(0.043)
sh	1.014***	0.699***	1.313***
	(0.029)	(0.072)	(0.028)
us	0.118	0.074	0.073
	(0.102)	(0.291)	(0.097)
Constant	1.001***	1.002***	1.001***
	(0.0005)	(0.001)	(0.0005)
Observations	3,143	605	3,207
R^2	0.298	0.171	0.419
Adjusted R^2	0.297	0.164	0.418
Residual Std. Error	0.028 (df = 3137)	0.033 (df = 599)	0.026 (df = 3201)
F Statistic	266.743*** (df = 5; 3137)	24.734*** (df = 5; 599)	462.315*** (df = 5; 3201)
Note:			*p<0.1; **p<0.05; ***p<0.01

The result shows that the us is not a significant feature at level 0.05. From the following scatter plots, we find that the covariates au, ag and pt have pairwise correlations and thus the covariates are not independent.

To check the multicolinearity for independent variables, we calculate their VIF. It seems that no necessarily multicolinearity among the covariates.

Since the estimation by LSE is not that satisfactory, we consider some basic variable selection method.

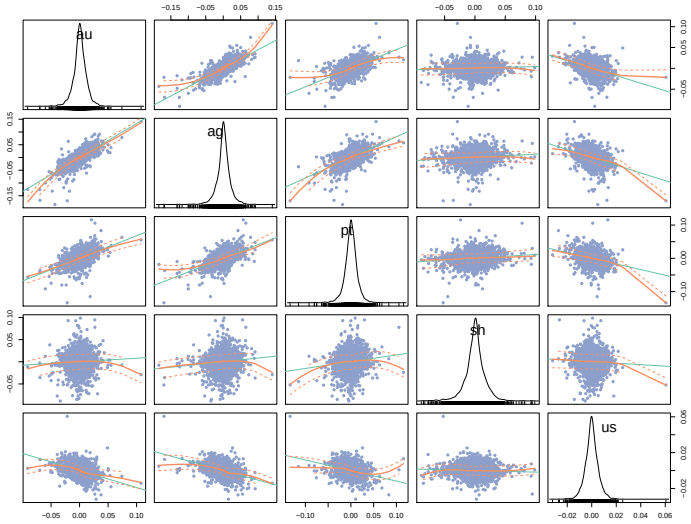
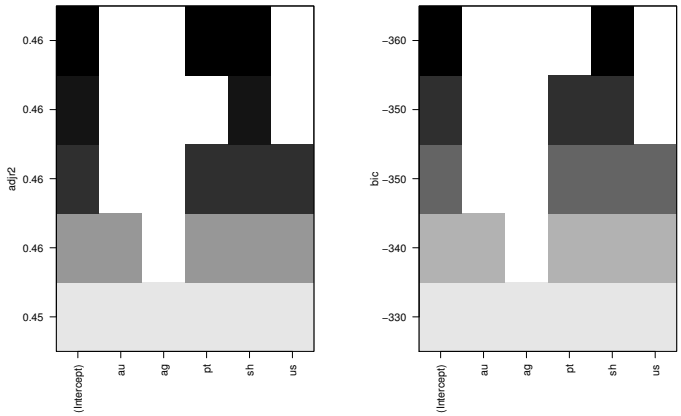


Figure 5: The pairwise plots of Au, Ag, Pt, US and SH

	au	ag	pt	sh	us
VIF	2.739	2.751	1.985	1.054	1.114

Table 4: VIF of variables.



	au	ag	pt	sh	us
Adj. R^2			√	√	
AIC				√	
BIC				√	
C_p				√	

Table 5: Summary of the variable selection results.

Finally, we conclude that the feature sh should be included in the linear regression model by the votes of theses variable selection methods, which is also an reasonable results referring to the scatter plots.

Exercise 2.1 ⋆⋆ Show that the sum of the diagonal leverages is $(p+1)/n$.

Exercise 2.2 ⋆⋆ Show that the adjusted $R^2 = 1 - (1 - R^2)(n-1)/(n-p)$, hence the adjust R^2 is always smaller than the R^2.

Exercise 2.3 ⋆⋆ Exercise 1 of RU Chapter 12.

Exercise 2.4 ⋆⋆ Exercise 5 of RU Chapter 12.

Exercise 2.5 ⋆⋆ Exercise 9 of RU Chapter 12.