Ch. 3: Linear Models of Regression

Chris Bishop’ s PRML Ch. 3: & ISL 3.2.3, 2.2.2; ESL 3.2, 7.3

王星

中国人民大学统计学院
E-mail:wangxingwisdom@126.com

November 1, 2020

Chapter content

- 一元线性回归;
- 多元回归
- 一致最优线性无偏估计BLUE
- 预测误差和均方误差（PE and MSE）
- 偏差与方差
- 模型选择和校验（Model selection and validation）
- 岭回归
- Lasso

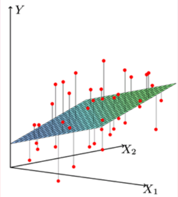
Linear Regression model

- 模型

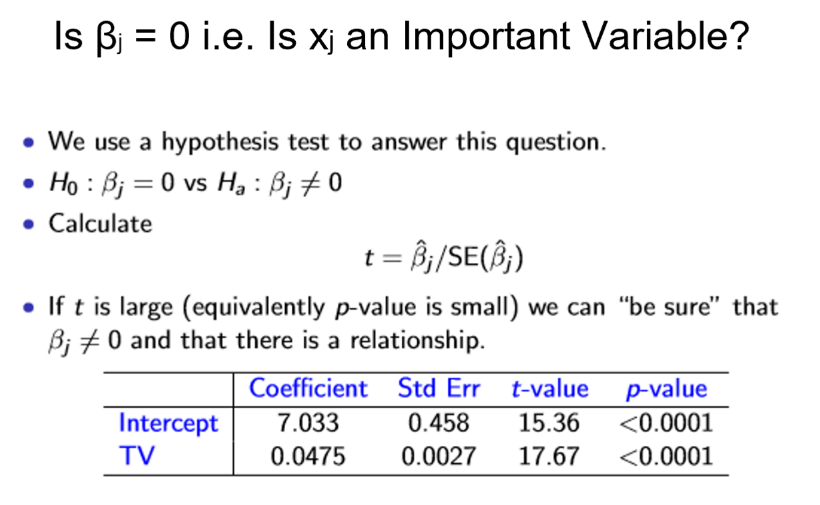
$$y = f(\mathrm{x}, \beta) = \beta_0 + \sum_{j=1}^{M-1} X_j \beta_j$$

- Here the X's might be
  - Raw predictor variables (continuous or coded-categorical);
  - Transformed predictors ($X_4 = \log X_3$)
  - Basis expansions ($X_4 = X_3^2, X_5 = X_3^3, etc.$)
  - Interactions ($X_4 = X_2 X_3.$)
- Popular choice for estimation is least squares:

$$RSS(\beta) = (Y - X\beta)^T (Y - X\beta)$$



Linear Regression coefficient



Is βj = 0 i.e. Is xj an Important Variable?

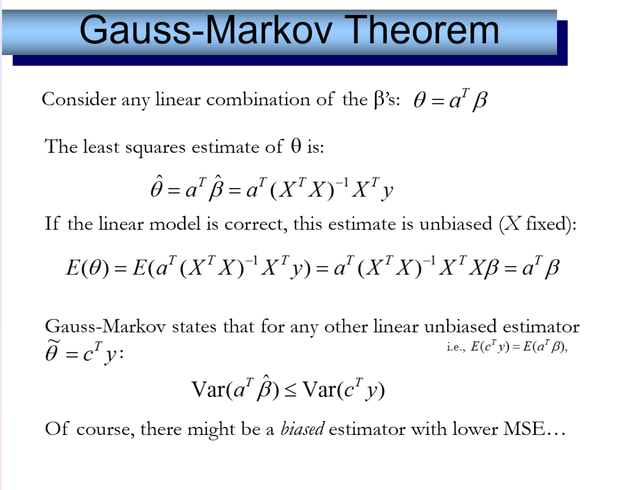
- We use a hypothesis test to answer this question.
- $H_0 : \beta_j = 0$ vs $H_a : \beta_j \neq 0$
- Calculate

$$t = \hat{\beta}_j / \mathrm{SE}(\hat{\beta}_j)$$

- If t is large (equivalently p-value is small) we can "be sure" that $\beta_j \neq 0$ and that there is a relationship.

	Coefficient	Std Err	t-value	p-value
Intercept	7.033	0.458	15.36	<0.0001
TV	0.0475	0.0027	17.67	<0.0001

Linear Regression model



Gauss-Markov Theorem

Consider any linear combination of the β's: $\theta = a^T \beta$

The least squares estimate of θ is:

$$\hat{\theta} = a^T \hat{\beta} = a^T (X^T X)^{-1} X^T y$$

If the linear model is correct, this estimate is unbiased (*X* fixed):

$$E(\theta) = E(a^T (X^T X)^{-1} X^T y) = a^T (X^T X)^{-1} X^T X\beta = a^T \beta$$

Gauss-Markov states that for any other linear unbiased estimator $\tilde{\theta} = c^T y$: i.e., $E(c^T y) = E(a^T \beta)$,

$$\mathrm{Var}(a^T \hat{\beta}) \leq \mathrm{Var}(c^T y)$$

Of course, there might be a *biased* estimator with lower MSE…

线性回归OLS估计的优点是无偏，在线性估计中方差最小，但在有可能存在有偏的估计但方差可能更小的估计。

Review: univariate regression1/2

Suppose that we have observations $y = (y_1, ..., y_n) \in R^n$, and we want to model these a linear function of $x = (x_1, ..., x_n) \in R^n$. The univariate linear regression coefficient of y on x is

$$\hat{\beta} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} = \frac{x^T y}{||x||_2^2}$$

This value $\hat{\beta} \in R$ is optimal in the least squares sense:

$$\hat{\beta} = \text{argmin}_\beta \sum_{i=1}^n (y_i - \beta x_i)^2 = \text{argmin}_\beta ||y - \beta x||_2^2.$$

We often think of the observations y as coming from the model

$$y = \beta^* x + \epsilon.$$

where $x \in R^n$ are fixed (nonrandom) measurements, $\beta^* \in R$ is some true coefficient, and $\epsilon = (\epsilon_1, ..., \epsilon_n) \in R^n$ are errors with $\text{E}[\epsilon_i] = 0, \text{Var}(\epsilon_i) = \sigma^2, \text{Cov}(\epsilon_i, \epsilon_j) = 0.$

一元回归系数估计及意义Review: univariate regression2/2

Now add an intercept term to the linear model:

$$y = \beta_0^* + \beta_1^* x + \epsilon$$

Estimate $\hat{\beta}_0, \hat{\beta}_1$ using least squares,

$$\hat{\beta}_0, \hat{\beta}_1 = \operatorname{argmin}_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 = \operatorname{argmin}_{\beta_0, \beta_1} \|y - \beta_0 \mathbb{1} - \beta_1 x\|$$

giving:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \quad \hat{\beta}_1 = \frac{(x - \bar{x}\mathbb{1})^T (y - \bar{y}\mathbb{1})}{\|x - \bar{x}\mathbb{1}\|_2^2}$$

Notice that

$$\hat{\beta}_1 = \frac{\mathrm{cov}(x, y)}{\mathrm{var}(x)} = \mathrm{cor}(x, y) \sqrt{\frac{\mathrm{var}(y)}{\mathrm{var}(x)}}$$

Review: multivariate regression1/2

- Now suppose that we are considering $y \in \mathbb{R}^n$ as a function of multiple predictors $X_1, \ldots, X_p \in \mathbb{R}^n$. We collect these predictors into columns of a predictor matrix $X \in \mathbb{R}^{n \times p}$. We assume that $X_1, \ldots, X_p$ are linearly independent($p \leq n$), so that $\text{rank}(X) = p$
- The model
$$y = X\beta^* + \epsilon;$$
where $X \in \mathbb{R}^{n \times p}$ is considered fixed, $\beta^* = (\beta_1^*, \ldots, \beta_p^*) \in \mathbb{R}^p$ are the true coefficients, and the errors $\epsilon = (\epsilon_1, \ldots, \epsilon_n) \in \mathbb{R}^n$ are as before (i.e., satisfying $\mathrm{E}[\epsilon] = 0$ and $\mathrm{Cov}(\epsilon) = \sigma^2 I$)
- For an intercept term, we can just append a column $\mathbb{1} \in \mathbb{R}^n$ of all 1s to the matrix X
- Estimate the coefficients $\hat{\beta} \in \mathbb{R}^p$ by least squares:
$$\hat{\beta} = \text{argmin}_{\beta \in R^p} \|y - X\hat{\beta}\|_2^2$$
- This gives
$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Geometry of Multivariate regression2/2

The fitted values are

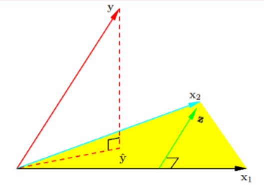
$$\hat{y} = X\hat{\beta} = X(X^T X)^{-1} X^T y$$

This is a linear function of y, $\hat{y} = Hy$, where $H = X(X^T X)^{-1} X^T$ is sometimes called the hat matrix.

The linear regression $\hat{y} \in \mathbb{R}^n$ is exactly the projection of $y \in \mathbb{R}^n$ onto the linear subspace $\text{span}\{X_1, \ldots, X_p\} = \text{col}(X) \subseteq R^n$.

projection matrix:

- The matrix H is symmetric: $H^T = H$;
- idempotent: $H^2 = H$;
- $Hx = x$ for all $x \in \text{col}(X)$, and $Hx = 0$ for all $x \perp L$.



Orthogonal complement

E.g., for any subspace $L \subseteq R^n$, its orthogonal complement is $L^\perp = \{x \in \mathbb{R}^n : x \in L\} = \{x \in \mathbb{R}^n : x \perp v \text{ for any } v \in L\}$.
Fact: $P_L + P_{L^\perp} = I$, so that $P_{L^\perp} = I - P_L$.

Hence for the linear regression of y on X, the residual vector is

$$y - \hat{y} = (I - P_{\mathrm{col}(X)})y = P^\perp_{\mathrm{col}(X)} y$$

So $y - \hat{y}$ is orthogonal to any $v \in \mathrm{col}(X)$; In particular, this means that $y - \hat{y}$ is orthogonal to each of $X_1, \ldots, X_p$.
E.g., the projection map P_L onto any linear subspace $L \in \mathbb{R}^n$ is always non-expansive, that is, for any points $x, z \in \mathbb{R}^n$,

$$\|P_L x - P_L z\|_2 \leq \|x - z\|_2;$$

Hence if $y_1, y_2 \in \mathbb{R}^n$ and $\hat{y}_1, \hat{y}_2 \in \mathbb{R}^n$ are their regression fits, then

$$\|\hat{y}_1 - \hat{y}_2\| = \|P_{\mathrm{col}(X)} y_1 - P_{\mathrm{col}(X)} y_2\|_2 \leq \|y_1 - y_2\|_2;$$

Best linear Best linear unbiased estimate (BLUE)

A natural question is: what is the best linear unbiased estimate (BLUE) $c^T y$ for estimating $a^T \beta^*$? Recall that the linear regression estimate $a^T \hat{\beta} = b^T y$ falls into this category (linear and unbiased) By "best" here, we mean the estimate $c^T y$ that minimizes the mean squared error in estimating $a^T \beta^*$:

$$\mathrm{MSE}(c^T y) = E[(c^T y - a^T \beta^*)^2];$$

Gauss-Markov theorem: the linear regression estimate $a^T \beta^{=} b^T y$is the BLUE, i.e., if $c^T y$ is any other unbiased estimate of $a^T \beta^*$, then

$$\mathrm{MSE}(a^T \hat{\beta}) \leq \mathrm{MSE}(c^T y);$$

Gauss-Markov theorem equivalently says that the regression estimate $a^T \hat{\beta}$ has smallest variance compared to all linear unbiased estimates

Estimate

Write $\langle a, b\rangle = a^T b = \sum_{i=1}^n a_i b_i$ as the inner-product for vectors $a, b \in \mathbb{R}^n$

In this notation, we can write the univariate linear regression coefficient of $y \in \mathbb{R}^n$ on a single predictor $x \in \mathbb{R}^n$ as

$$\hat{\beta} = \frac{\langle x, y\rangle}{\|x\|_2^2}$$

Given p predictor variables $X_1, \ldots, X_p \in \mathbb{R}^n$, the univariate linear regression coefficient of y on X_j is

$$\hat{\beta}_j = \frac{\langle X_j, y\rangle}{\|X_j\|_2^2}.$$

Fact: if $X_1, \ldots, X_p$ are orthogonal, then this is also the coefficient of X_j in the multivariate linear regression of y on all of $X_1, \ldots, X_p$.

Univariate regression with intercept

For univariate linear regression with an intercept term, i.e., for regressing $y \in \mathbb{R}^n$ on predictors $\mathbb{1}, x \in \mathbb{R}^n$,we can write the coefficient of x as

$$\hat{\beta}_1 = \frac{\langle x - \bar{x}\mathbb{1}, y \rangle}{\|x - \bar{x}\mathbb{1}\|_2^2}$$

We can alternatively view this as result of two steps:

- Regress x on $\mathbb{1}$, yielding the coefficient

$$\frac{\langle \mathbb{1}, x \rangle}{\|\mathbb{1}\|_2^2} = \frac{\langle \mathbb{1}, x \rangle}{n} = \bar{x}$$

  and the residual $z = x - \bar{x}\mathbb{1} \in \mathbb{R}^n$

- Regress y on z, yielding the coefficient

$$\hat{\beta}_1 = \frac{\langle z, y \rangle}{\|z\|_2^2} = \frac{\langle x - \bar{x}\mathbb{1}, y \rangle}{\|x - \bar{x}\mathbb{1}\|_2^2}$$

Multivariate regression by orthogonalization 1/3(successive orthogonalization))

This idea extends to multivariate linear regression of $y \in \mathbb{R}^n$ on predictors $X_1, \ldots, X_p \in R^n$. Consider the p-step procedure:

1. Let $Z_1 = X_1$;
2. For $j = 2, \ldots, p$: Regress X_j onto $Z_1, \ldots, Z_{j-1}$ to get coefficients $\hat{\gamma}_{jk} = \frac{\langle Z_k, X_j \rangle}{\|Z_k\|_2^2}$ for $k = 1, \ldots, j-1$, and residual vector

$$Z_j = X_j - \sum_{k=1}^{j-1} \hat{\gamma}_{jk} Z_k$$

3. Regress y on Z_p to get the coefficient $\hat{\beta}_p$.

Multivariate regression by orthogonalization(2/3)

1. The vectors $Z_1, \ldots, Z_p \in \mathbb{R}^n$ produced by this algorithm are orthogonal.
2. For any $j = 1, \ldots, p$, the definition $Z_j = X_j - \sum_{k=1}^{j-1} \gamma_{jk} Z_k$ shows that each Z_j is a linear combination of $X_1, \ldots, X_j$, In fact, $\operatorname{span}\{X_1, \ldots, X_j\} = \operatorname{span}\{Z_1, \ldots, Z_j\}$.
3. the linear regression fit y on $X_1, \ldots, X_p$ is the same as the linear regression fit of y on $Z_1, \ldots, Z_p$. Call this fit $\hat{y}$.

   $$y = c_1 Z_1 + \ldots + c_p Z_p;$$

   for some $c_1, \ldots, c_p$
4. As $Z_1, \ldots, Z_p$ are orthogonal, the coefficients $c_1, \ldots, c_p$ are just given by univariate linear regression, so in particular we have

   $$c_p = \frac{\langle Z_p, y \rangle}{\|Z_p\|_2^2} = \hat{\beta}_p.$$
5. For each Z_j in the expression

   $$\hat{y} = c_1 Z_1 + \ldots + c_p Z_p + \hat{\beta}_p Z_p$$

Multivariate regression by orthogonalization 3/3

plug in the linear representation in terms of $X_1, \ldots, X_p$. Note that the variable X_p appears only through Z_p, and the coefficient of X_p is $\mathbb{1} : Z_p = X_p - \sum_{k=1}^{p-1} \hat{\gamma}_{pk} Z_k$.

Claim

the output $\hat{\beta}_p$ of this algorithm is exactly the coefficient of X_p in the multivariate linear regression of y on $X_1, \ldots, X_p$.

$$\hat{\beta}_p = \frac{\langle Z_p, y \rangle}{\|Z_p\|_2^2}, \quad \hat{\beta}_j = \frac{\langle Z_j, y \rangle}{\|Z_j\|_2^2}$$

where Z_p the residual from regressing X_p onto $Z_1, \ldots, Z_{p-1}$, i.e., the residual from regressing X_p onto $X_1, \ldots, X_{p-1}$.

Orthogonal and correlated predictor variables

1. If $X_1, \ldots, X_p$ are orthogonal, then we claimed last slides that the jth multiple regression coefficient of y on $X_1, \ldots, X_p$ is equal to the univariate regression coefficient of y on X_j.
2. If $X_1, \ldots, X_p$ are correlated, Note that z_j is the residual from regressing X_j onto $X_i, i \neq j$. Remember that the regression fit of X_j onto $X_i, i \neq j$ is really just the projection of X_j onto the linear subspace $\text{span}\{X_i : i \neq j\}$.
3. If X_j is highly correlated with the rest, then this fit is close to X_j, so the residual z_j is close to 0. This makes the regression coefficient $\hat{\beta}_j = \frac{\langle z_j, y \rangle}{\|z_j\|_2^2}$ unstable, as the denominator is very small, but the numerator can be too.

Variance inflation

From this formula we can explicitly compute the variance of the *j*th multiple regression coefficient:

$$\mathrm{Var}(\hat{\beta}_j) = \frac{\mathrm{Var}(\langle z_j, y\rangle)}{\|z_j\|_2^4} = \frac{\|z_j\|_2^2\sigma^2}{\|z_j\|_2^4} = \frac{\sigma^2}{\|z_j\|_2^2}$$

Having correlated predictors inflates the variance of multiple regression coefficients. Remember that the Z-statistic for the *j*th regression coefficient is

$$Z_j = \frac{\hat{\beta}_j}{\sqrt{\mathrm{Var}(\hat{\beta}_j)}} = \frac{\hat{\beta}_j}{\sigma} \cdot \|z_j\|_2$$

so if X_j is highly correlated with the other predictors, its regression coefficient will likely be not significant (according to Z_j)

Dropping predictor variables

- Now suppose that X_j and X_k both contribute in explaining y, but are highly correlated with each other. Then from what we said on the last slide, neither $||Z_j||$ nor $||Z_k||$ will be very large, so they won't be significant.
- Now what happens if we remove one of them—say, X_k—from the model, and recompute the regression coefficients? The term $||z_j||_2^2$ will be much larger (assuming that X_j is not highly correlated with other predictors than X_k). Hence it's variance will decrease, and Z_j will likely increase,;
- This is why we can't remove two (or more) supposedly insignificant predictors at a time—in short: because significance depends on what other predictors are in the model!

Shortcomings of regression

Two main themes:

- Predictive ability: the linear regression often does not predict well, especially when p (the number of predictors)is large (Important to note that is not even necessarily due to nonlinearity in the data, Can still predict poorly even when a linear model could not well);
- Interpretative ability: linear regression ”freely” assigns a coefficient to each predictor variable. When p is large, we may sometimes seek, for the sake of interpretation, a smaller set of important variables;

Hence we want to ”encourage” fitting procedure to make only a subset of the coefficients large, and others small or even better, zero

Prediction accuracy and mean-squared error

Suppose we observe data of the form

$$y_i = f(x_i) + \epsilon_i, i = 1, \ldots, n$$

Here $f : \mathbb{R}^p \to \mathbb{R}$ is some true function, $x_i = (x_{i1}, \ldots, x_{ip}) \in \mathbb{R}^p$ are fixed predictor measurements, and $\epsilon_i \in R^n$ are random errors with $E[\epsilon_i] = 0$, $\mathrm{Var}(\epsilon_i) = \sigma^2$, and $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0$.
Consider one more data point y_0, independent of $y_1, \ldots, y_n$,

$$y_0 = f(x_0) + \epsilon_0;$$

and suppose that we want to predict y_0 at the fixed point $x_0 \in \mathbb{R}^p$, from the observed pairs $(y_1, x_1), \ldots, (y_n, x_n)$.Think of, e.g., the typical linear regression model: here we have $f(x_i) = x_i^T \beta^*$, for some true regression coefficients β^* Suppose that we use $\hat{f}$ to predict f (again, think of regression: $\hat{f}(x_i) = x_i^T \hat{\beta}$). In particular, we predict y_0 via $\hat{f}(x_0)$. Prediction error is

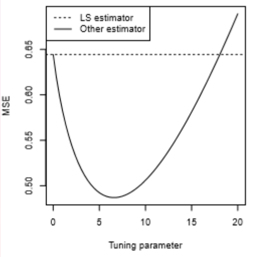
$$\begin{aligned}\mathrm{PE}(\hat{f}(x_0)) =& \mathrm{E}[(y_0 - \hat{f}(x_0))^2]) = \mathrm{E}[(y_0 - f(x_0))^2] + \mathrm{E}[(f(x_0) - \hat{f}(x_0))^2] \\ =& \sigma^2 + \mathrm{MSE}(\hat{f}(x_0)) = \sigma^2 + [\mathrm{Bias}\, \hat{f}(x_0))]^2 + [\mathrm{Var}\, \hat{f}(x_0))]\end{aligned}$$

Example: small regression coefficients

Recall the Gauss-Markov theorem said that this estimator is the BLUE: best linear unbiased estimator. i.e., for a fixed input point x_0, if $\hat{f}(x_0)$ is any other linear, unbiased estimator of $x_0^T\beta^*$, then

$$\mathrm{MSE}(\hat{f}(x_0)) \geq \mathrm{MSE}(f^{\mathrm{LS}}(x_0)) = \mathrm{MSE}(x_0^T\hat{\beta})$$

- Unbiased: this means that $\mathrm{E}[\hat{f}(x_0)] = x_0^T\beta^*$;
- Linear: this means linear in $y = (y_1, \ldots, y_n), i.e., \hat{f}(x_0) = c^T y$ for some c.



Averaging over all inputs

Average PE or MSE across all the input points $x_1, \dots, x_n$

$$\mathrm{PE}(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{PE}(\hat{f}(x_i)), \quad \mathrm{MSE}(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{MSE}(\hat{f}(x_i))$$

Note the same relationships hold:

$$\mathrm{PE}(\hat{f}) = \sigma^2 + \mathrm{MSE}(\hat{f}) = \sigma^2 + \frac{1}{n}\sum_{i=1}^{n}[\mathrm{Bias}(\hat{f}(x_i))]^2 + \frac{1}{n}\sum_{i=1}^{n}[\mathrm{Var}(\hat{f}(x_i))].$$

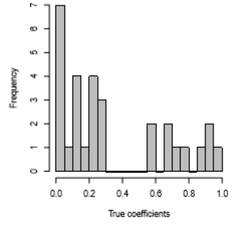
We set $\hat{f}^{LS}(x_i) = x_i^T\hat{\beta}$

$$\begin{aligned}\mathrm{PE}(\hat{f}^{LS}) &= \sigma^2 + \frac{1}{n}\sum_{i=1}^{n}[\mathrm{Bias}(x_i^T\hat{\beta})]^2 + \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}(x_i^T\hat{\beta}) \\ &= \sigma^2 + 0 + \frac{p\sigma^2}{n}\end{aligned}$$

This scales linearly with the number of predictors p

Example: small regression coefficients1/2

Example: simulation with $n = 50$ and $p = 30$. The entries of the predictor matrix $X \in \mathbb{R}^{50\times 30}$ are all i.i.d. $N(0,1)$, so overall the variables have low correlation Histogram of the true regression coefficients $\beta^* \in \mathbb{R}^{30}$:



Here 10 coefficients are large(between 0.5 and 1) and 20 coefficients are small (between 0 and 0.3)

Example: small regression coefficients2/2

The response $y \in \mathbb{R}^{50}$ is drawn from the model $y = X\beta^* + \epsilon$, where the entries of $\epsilon \in \mathbb{R}^{50}$ are are all i.i.d $N(0,1)$ (hence the noise variance is $\sigma^2 = 1$.)

We repeated the following 100 times:

- Generate a response vector y;
- Compute the linear regression fit $X\hat{\beta}$;
- Generate a new response y'
- Record the error $1/n \sum_{i=1}^{n} (y'_i - x_i^T \hat{\beta})^2$

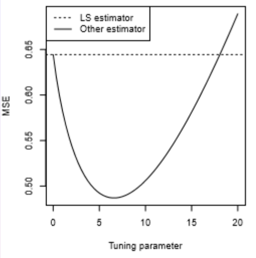
We averaged this observed error over the 100 repetitions to get an estimate of the the prediction error.

We also estimated the squared bias and variance of the fits $X\hat{\beta}$ over the 100 repetitions. Recall that it should be true that prediction error = 1 + squared bias+variance.

```
    > bias    var    p/n,  1+ bias + var,   prederr
[1] 0.00647163  0.6273129  0.6  1.633785  1.64436
```

How can we do better?

- For linear regression, its prediction error is just $\sigma^2 + p/n \cdot \sigma^2$, the second term being the variance $1/n \sum_{i=1}^{n} \mathrm{Var}(x_i^T \hat{\beta})$.
- What can we see from this? Each additional predictor variable will add the same amount of variance σ^2/n, regardless of whether its true coefficient is large or small (or zero).
- In the previous example, we were "spending" variance in trying to fit truly small coefficients—there were 20 of them, out of 30 total.
- So can we do better by shrinking small coefficients towards zero, incurring some bias, so as to reduce the variance? You can think of this as trying to ignore some "small details" in order to get a more stable "big picture."



Linear regression

Squared bias ≈ 0.006

Variance ≈ 0.627

Pred error $\approx 1{+}0.006{+}0.627$

≈ 1.633

Ridge regression at its best

Squared bias ≈ 0.077

Variance ≈ 0.403

Pred error $\approx 1{+}0.077{+}0.403$

≈ 1.48

Ridge regression

Ridge regression is like least squares but shrinks the estimated coefficients towards zero. Given a response vector $y \in \mathbb{R}^n$ and a predictor matrix $X \in \mathbb{R}^{n\times p}$, the ridge regression coefficients are defined as

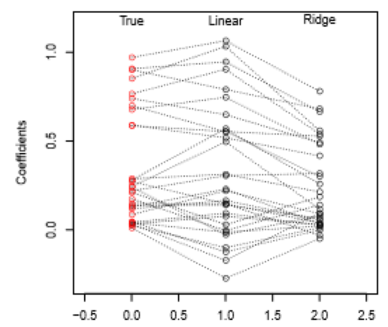
$$\begin{aligned}\hat{\beta}^{ridge} &= \operatorname{argmin}_{\beta\in\mathbb{R}^p} \sum_{i=1}^{n}(y_i - x_i^T\beta)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 \\ &= \operatorname{argmin}_{\beta\in\mathbb{R}^p} \underbrace{\|y - X\beta\|_2^2}_{} + \lambda \underbrace{\|\beta\|_2^2}_{}\end{aligned}$$

Here $\lambda \geq 0$ is a tuning parameter, which controls the strength of the penalty term. Note that:

- When $\lambda = 0$, we get the linear regression estimate;
- When $\lambda = \infty$, we get $\hat{\beta}^{ridge} = 0$;
- For λ in between, we are balancing two ideas: fitting a linear model of y on X, and shrinking the coefficients

visual representation of ridge coefficient

Recall last example ($n = 50$, $p = 30$, and $\sigma^2 = 1$; 10 large true coefficients, 20 small). Here is a visual representation of the ridge regression coefficients for $\lambda = 25$:



Important details

- When including an intercept term in the regression, we usually leave this coefficient unpenalized. Otherwise we could add some constant amount c to the vector y, and this would not result in the same solution. Hence ridge regression with intercept solves

$$\hat{\beta}_0, \hat{\beta}^{ridge} = \text{argmin}_{\beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^p} \|y - \beta_0 \mathbb{1} - X\beta\|_2^2 + \lambda \|\beta\|_2^2$$

- If we center the columns of X, then the intercept estimate ends up just being $\hat{\beta}_0 = \hat{y}$, so we usually just assume that y, X have been centered and don't include an intercept
- Also, the penalty term $\|\beta\|_2^2 = \sum_{j=1}^{p} \beta_j^2$ is unfair is the predictor variables are not on the same scale. (Why?) Therefore, if we know that the variables are not measured in the same units, we typically scale the columns of X (to have sample variance 1), and then we perform ridge regression

Bias and variance of ridge regression

The bias and variance are not quite as simple to write down for ridge regression as they were for linear regression, but closed-form expressions are still possible Recall that

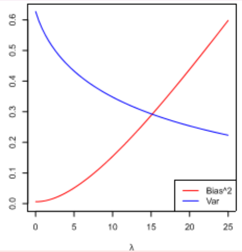
$$\hat{\beta}^{ridge} = \operatorname{argmin}_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2$$

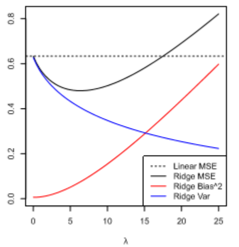
The general trend is:

- The bias increases as λ (amount of shrinkage) increases;
- The variance decreases as λ (amount of shrinkage) increases

What is the bias at $\lambda = 0$? The variance at $\lambda = \infty$?

$n = 50, p = 30, \sigma^2 = 1$; 10 large true coefficients, 20 small





Moderately regression coefficients

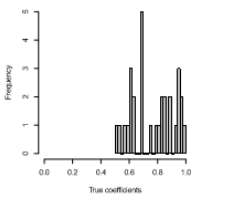
If all the true coefficients are moderately large, is it still helpful to shrink the coefficient estimates?

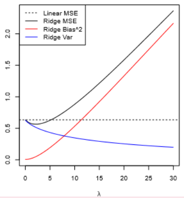
Moderately regression coefficients

If all the true coefficients are moderately large, is it still helpful to shrink the coefficient estimates?

The answer is (perhaps surprisingly) still ”yes”. But the advantage of ridge regression here is less dramatic, and the corresponding range for good values of λ is smaller.

- Same setup as last example: $n = 50$, $p = 30$, and $\sigma^2 = 1$. Except now the true coefficients are all moderately large (between 0.5 and 1). Histogram:





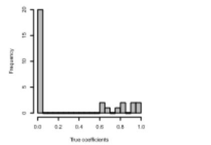
- The linear regression fit:Squared bias ≈ 0.006 Variance ≈ 0.628 Pred. error $\approx 1+0.006+0.628 \approx 1.634$.
- Only works for λ less than 5, otherwise it is very biasd.

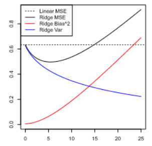
Variable selection

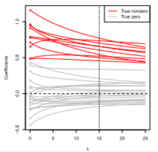
- Why are these numbers essentially the same as those from the last example, even though the true coefficients changed
- Ridge regression can still outperform linear regression in terms of mean squared error.
- To the other extreme (of a subset of small coefficients), suppose that there is a group of true coefficients that are identically zero. This means that the mean response doesn’ t depend on these predictors at all; they are completely extraneous.
- The problem of picking out the relevant variables from a larger set is called variable selection. In the linear model setting, this means estimating some coefficients to be exactly zero. Aside from predictive accuracy, this can be very important for the purposes of model interpretation

Subset of zero coefficients

- Same general setup as the running example: $n = 50, p = 30$, and $\sigma^2 = 1$. Now, the true coefficients: 10 are large (between 0.5 and 1) and 20 are exactly 0. Histogram:
- The linear regression fit;
- Ridge regression performs well in terms of mean-squared error;



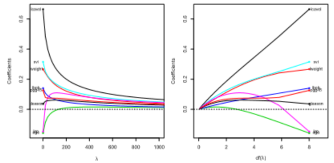




- Squared bias $\approx$ 0.006 Variance $\approx$ 0.627 Pred. error $\approx$ 1+0.006+0.627 $\approx$ 1.633
- The red paths correspond to the true nonzero coefficients;the gray paths correspond to true zeros. The vertical dashed line at $\lambda = 15$ marks the point above which ridge regression's MSE starts losing to that of linear regression Notice: The gray coefficient paths are not exactly zero; they are shrunken, but still nonzero.

Ridge regression doesn’t perform variable selection

- Prostate Data Example: the problem is interested in the level of prostate-specific antigen (PSA), elevated in men who have prostate cancer. Measurements of PSA on $n = 97$ men with prostate cancer, and $p = 8$ clinical predictors. Ridge coefficients: (after centering and scaling). The resulting coefficient profiles:



- Ridge regression doesn’t set coefficients exactly to zero unless $\lambda = +\infty$, in which case they’re all zero. Hence ridge regression cannot perform variable selection, and even though it performs well in terms of prediction accuracy, it does poorly in terms of offering a clear interpretation.
- This doesn’t give us a clear answer to the question ...
- Perform ridge regression over a wide range of λ values

Ridge regression

- Ridge regression, which minimizes the usual regression criterion plus a penalty term on the squared l_2 norm of the coefficient vector. As such, it shrinks the coefficients towards zero. This introduces some bias, but can greatly reduce the variance, resulting in a better mean-squared error.
- The amount of shrinkage is controlled by λ, the tuning parameter that multiplies the ridge penalty. Large λ means more shrinkage, and so we get different coefficient estimates for different values of λ. Choosing an appropriate value of λ is important, and also difficult.
- Ridge regression performs particularly well when there is a subset of true coefficients that are small or even zero. It doesn't do as well when all of the true coefficients are moderately large; however, in this case it can still outperform linear regression over a pretty narrow range of (small) λ values.

The lasso: Least Absolute Selection and Shrinkage Operator

The lasso estimate is defined as

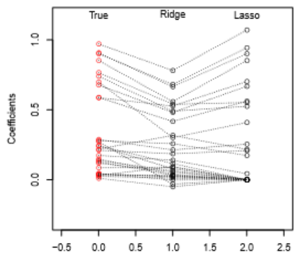
$$\begin{aligned}\hat{\beta}^{lasso} &= \mathrm{argmin}_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \\ &= \mathrm{argmin}_{\beta \in \mathbb{R}^p} \underbrace{\|y - X\beta\|_2^2}_{} + \lambda \underbrace{\|\beta\|}_{}\end{aligned}$$

The only difference between the lasso problem and ridge regression is that the latter uses a (squared) l_2 penalty $\|\beta\|_2^2$, while the former uses an l_1 penalty $\|\beta\|_1$. But even though these problems look similar, their solutions behave very differently.

- When $\lambda = 0$, we get the linear regression estimate; When $\lambda = \infty$, we get $\hat{\beta}^{lasso} = 0$;
- For λ in between, we are balancing two ideas: fitting a linear model of y on X, and shrinking the coefficients,But the nature of the l_1 penalty causes some coefficients to be shrunken to zero exactly.

The lasso: Least Absolute Selection and Shrinkage Operator

- This is what makes the lasso substantially different from ridge regression: it is able to perform variable selection in the linear model. As λ increases, more coefficients are set to zero (less variables are selected), and among the nonzero coefficients, more shrinkage is employed.
- Last running example from last time: $n = 50, p = 30, \sigma^2 = 1$, 10 large true coefficients, 20 small. Here is a visual representation of lasso vs. ridge coefficients (with the same degrees of freedom);



Lasso regression coefficients with intercept and scales

- How to do with intercept? When including an intercept term in the model, we usually leave this coefficient unpenalized, just as done with ridge regression. Lasso problem with intercept is

$$\hat{\beta}_0, \hat{\beta}^{lasso} = \text{argmin}_{\beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^p} \|y - \beta_0 \mathbb{1} - X\beta\|_2^2 + \lambda \|\beta\|$$

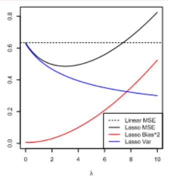
- As seen before, if the columns of X are centered, then the intercept estimate turns out to be $\hat{\beta}_0 = \bar{y}$. Therefore we typically center y, X and don't include an intercept.
- As with ridge regression, the penalty term $\|\beta\|_1 = \sum_{j=1}^{p} \|\beta_j\|$ is not fair is the predictor variables are not on the same scale. Hence, if it's known that the variables are not on the same scale to begin with, we scale the columns of X(to have sample variance 1), and then we solve the lasso problem.

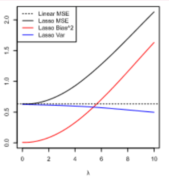
Bias and variance of the lasso

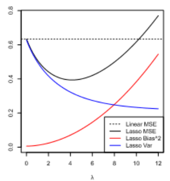
Generally speaking, the bias increases as λ (amount of shrinkage) increases and the variance decreases as λ (amount of shrinkage) increases.

What is the bias at $\lambda = 0$? The variance at $\lambda = \infty$?In terms of prediction error (or mean squared error), the lasso performs comparably to ridge regression(Use package lars)

- For subset of small coefficients example(left): $n = 50, p = 30$; true coefficients: 10 large, 20 small;
- For subset of moderate coefficients example(mid): $n = 50, p = 30$; true coefficients: 30 moderately large(Note that here, as opposed to ridge regression the variance doesn't decrease fast enough to make the lasso favorable for small λ)
- For subset of zero coefficients example(right): $n = 50, p = 30$; true coefficients: 10 large, 20 zero

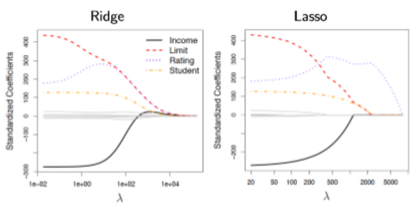






Example: credit data

Example from ISL sections 6.6.1 and 6.6.2: response is average credit debt, predictors are income, limit (credit limit), rating (credit rating), student (indicator), and others



Constrained form

$$\hat{\beta}^{ridge} = \text{argmin}_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 \text{ subject to} \|\beta\|_2^2 \leq t$$

$$\hat{\beta}^{lasso} = \text{argmin}_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 \text{ subject to} \|\beta\|_1 \leq t$$

- Now t is the tuning parameter (before it was λ). For any λ and corresponding solution in the previous formulation (sometimes called penalized form), there is a value of t such that the above constrained form has this same solution.
- In comparison, the usual linear regression estimate solves the unconstrained least squares problem; these estimates constrain the coefficient vector to lie in some geometric shape centered around the origin. This generally reduces the variance because it keeps the estimate close to zero. But which shape to choose really matters!

What is degrees of freedom 1/2

Broadly speaking, the degrees of freedom of an estimate describes its effective number of parameters

- More precisely, given data $y \in \mathbb{R}^n$ from the model

  $$y_i = \mu_i + \epsilon_i, i = 1, \ldots, n.$$

  Where $\mathrm{E}(\epsilon_i) = 0, \mathrm{Var}(\epsilon_i) = \sigma^2, \mathrm{Cov}(\epsilon_i, \epsilon_j) = 0$, suppose that we estimate y by $\hat{y}$. The degrees of freedom of the estimate $\hat{y}$ is

  $$\mathrm{df}(\hat{y}) = \frac{1}{\sigma^2} \sum_{i=1}^{n} \mathrm{Cov}(\hat{y}_i, y_i)$$

- The higher the correlation between the ith fitted value and the ith data point, the more adaptive the estimate, and so the higher its degrees of freedom.

What is degrees of freedom 2/2

Let $X \in \mathbb{R}^{n\times p}$ be a fixed matrix of predictors

- For linear regression, $\hat{y} = X\hat{\beta}^{linear}$, df$(\hat{y}) = p$.
- For ridge regression,$\hat{y} = X\hat{\beta}^{ridge}$, df$(\hat{y})$ = trace$((X^T X + \lambda I)^{-1} X^T)$.
- For thelasso, $\hat{y} = X\hat{\beta}^{lasso}$, df$(\hat{y})$= E[number of nonzero coefficients in β^{lasso}.]

作业(2019/10/17)

- 从R软件包中的ISLR数据包中提取credit数据，balance是目标变量，其他定量变量是输入变量，进行回归建模，比较以下三种模型的效果
  - 请输出每个变量对balance的一元线性回归，讨论回归系数的相关系数之间的关系；
  - 请输出普通线性回归多元回归系数，讨论它和一元线性回归系数之间的差异，分析这种差异是怎样产生的；
  - 请根据successive orthogonalization重新计算回归系数，讨论(3)和(2)之间的差异；
  - 请尝试ridge回归，设置不同的λ输出系数，讨论(2)(3)(4)的模型拟合MSE，讨论以上四种系数的异同.
- 将slides第23页的模拟重新做一遍，不改变有贡献的变量的个数，其真实的系数在$0.5 \sim 1$之间，冗余变量的系数在$-0.5 \sim 0.5$之间，观察prediction error 和偏差，模型方差之间的关系，比较ridge 与普通线性回归之间的PE。

作业(2020/10/29)

从R软件包中的ISLR数据包中提取credit数据，balance是目标变量，其他定量变量是输入变量，进行回归建模，比较以下几种模型的效果

- 请输出balance对每个定量自变量的一元线性回归，讨论回归系数与相关系数之间的关系；
- 请输出普通线性回归(OLS)的多元回归系数，讨论它和一元线性回归系数之间的差异，分析这些差异是怎样产生的；
- 请尝试successive orthogonalization系数估计法，该系数估计过程如果越到残差几乎为0的情况，请绕过该变量，只讲残差不为0的变量引入模型，估计回归系数，比较和多元回归系数的结果差异
- 请尝试ridge回归，设置不同的λ输出系数，讨论(2)(3)模型拟合的MSE；
- 请尝试lasso回归，设置不同的λ输出系数，讨论(2)(3)模型拟合的MSE；
- 请根据lasso回归选择的变量继续尝试successive orthogonalization，重新计算回归系数，讨论(3)-(6) 之间的差异；讨论以上四种系数的异同.