$$E(Y-\hat{Y})^2 = E[f(X)+\epsilon-\hat{f}(X)]^2 = \underbrace{[f(X)-\hat{f}(X)]^2}_{\text{Reducible}} + \underbrace{\text{Var}(\epsilon)}_{\text{Irreducible}},$$

standard error of μˆ, written as SE(μˆ), σ is the standard deviation of each of the realizations yi of Y

$$\text{Var}(\hat{\mu}) = \text{SE}(\hat{\mu})^2 = \frac{\sigma^2}{n} \quad \text{SE}(\hat{\beta}_0)^2 = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right] \quad \text{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i-\bar{x})^2}$$

<expected test MSE at X0>

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \text{Var}(\hat{f}(x_0)) + [\text{Bias}(\hat{f}(x_0))]^2 + \text{Var}(\epsilon).$$

- σ is known as the Residual Standard Error

$$\text{RSE} = \sqrt{\text{RSS}/(n-2)}.$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

Confidence Interval

$$[\hat{\beta_i} - t_{n-p-1,\alpha/2}.\text{SE}(\hat{\beta_i}), \hat{\beta_i} + t_{n-p-1,\alpha/2}.\hat{\text{SE}}(\beta_i)]$$

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$$
$$Var(cX) = c^2 Var(X)$$
$$Cov(cX, Y) = Cov(X, cY) = cCov(X, Y)$$

$$\text{Cor}(X,Y) = \frac{\sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^n (x_i-\bar{x})^2}\sqrt{\sum_{i=1}^n (y_i-\bar{y})^2}}$$

$$\text{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2$$
$$= (y_1-\hat{\beta}_0-\hat{\beta}_1 x_1)^2 + (y_2-\hat{\beta}_0-\hat{\beta}_1 x_2)^2 + \cdots + (y_n-\hat{\beta}_0-\hat{\beta}_1 x_n)^2.$$

* The least squares approach chooses βˆ0 and βˆ1 to minimize the RSS

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^n (x_i-\bar{x})^2} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

Variance: the amount by which fˆ would change if we estimated it using a different training data set.

Bias: the error that is introduced by approximating a real-life problem.

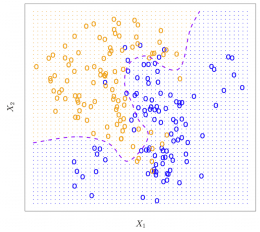
As a general rule, as we use more flexible methods, the variance will increase and the bias will decrease.

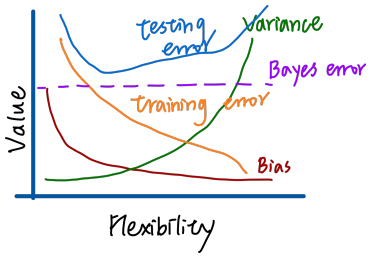
Classification

The Bayes Classifier: assigns each observation to the most likely class, given its predictor values.

$$\Pr(Y = j|X = x_0)$$

1) predicting class one if Pr(Y = 1|X = x0) > 0.5, and class two otherwise.
2) The Bayes classifier produces the lowest possible test error rate, called the Bayes error rate.



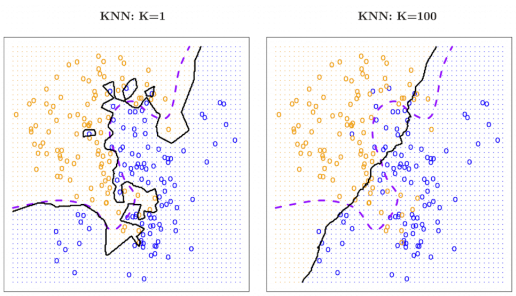


3) Bayes (irreducible) error - defines the lower limit, the test error is bounded below by the irreducible error due to variance in the error (epsilon) in the output values (0 <= value). When the training error is lower than the irreducible error, overfitting has taken place.

K-Nearest Neighbors:

K = 1, low bias but very high variance

As K grows, the method becomes less flexible and produces a decision boundary that is close to linear.



<Accessing the accuracy of model>

1) ***Residual Standard Error(RSE):*** an estimate of the standard deviation of ε, the average amount that the response will deviate from the true regression line(lack of fit of the model).

$$\text{RSE} = \sqrt{\frac{1}{n-2}\text{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}.$$

2) ***R^2 Statistic:*** a measure of the linear relationship between X and Y, the proportion of variance explained, between 0 & 1.

$$R^2 = \frac{\text{TSS}-\text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

$\text{TSS} = \sum(y_i - \bar{y})^2$ is the *total sum of squares*

(in the simple linear regression setting, R^2 = r^2=(Cor (X,Y)^2)

Linear Regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon,$$

We interpret βj as the average effect on Y of a one unit increase in Xj, holding all other predictors fixed.

To test the null hypothesis, we need to determine whether βˆ1, our estimate for β1, is sufficiently far from zero that we can be confident that β1 is non-zero.

$$t = \frac{\hat{\beta}_1 - 0}{\text{SE}(\hat{\beta}_1)}$$

measures the number of standard deviations that βˆ1 is away from 0

<Multiple Linear Regression>

1) Is There a Relationship Between the Response and Predictors?

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0 \qquad H_a : \text{ at least one } \beta_j \text{ is non-zero.}$$

$$F = \frac{(\text{TSS}-\text{RSS})/p}{\text{RSS}/(n-p-1)}$$

If no relationship between the response and predictors : F=1.
If Ha is true, then E{(TSS − RSS)/p} > σ2, so we expect F to be greater

(a larger F is needed to reject H0 if n is small)

Based on this p-value, we can determine whether or not to reject H0

$$F = \frac{(\text{RSS}_0 - \text{RSS})/q}{\text{RSS}/(n-p-1)}$$

if particular subset q of the coefficients are zero. we fit a second model that uses all the variables except those last q. residual sum of squares for that model is RSS0.

2) Deciding on Important Variables

Variable Selection: Forward / Backward / Mixed

Quality of a model: Mallow's Cp/ AIC/ BIC/ adjusted R^2

3) Model Fit: RSE/ R^2

$$\text{RSE} = \sqrt{\frac{1}{n-p-1}\text{RSS}}$$

models with more variables can have higher RSE if the decrease in RSS is small relative to the increase in p.

4) Prediction

The Validation Set Approach: randomly dividing the available set of observations into two parts, a training set and a validation set or hold-out set.

Drawbacks:

1) the validation estimate of the test error rate can be highly variable

2) In the validation approach, only a subset of the observations, validation set error rate may tend to overestimate the test error rate

Cross-Validation on Classification: rather than using MSE to quantify test error, we instead use the number of misclassified observations

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{Err}_i \quad \text{where } \mathrm{Err}_i = I(y_i \neq \hat{y}_i)$$

Leave-One-Out Cross-Validation: a single observation (x1,y1) is used for the validation set, and the remaining observations {(x2, y2), . . . , (xn, yn)} make up the training set.

Properties:

1) unbiased estimate for the test error
2) highly variable, since it is based upon a single observation
3) always yield the same results
4) $\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{MSE}_i$

Bootstrap: allows us to use a computer to emulate the process of obtaining new sample sets. We obtain distinct data sets by repeatedly sampling observations from the original data set. The sampling is performed with replacement, which means that the with same observation can occur more than once in the bootstrap data set.

ex. repeated B times, compute the standard error of these bootstrap

$$\mathrm{SE}_B(\hat{\alpha}) = \sqrt{\frac{1}{B-1}\sum_{r=1}^{B}\left(\hat{\alpha}^{*r} - \frac{1}{B}\sum_{r'=1}^{B}\hat{\alpha}^{*r'}\right)^2}.$$

k-Fold Cross-Validation: randomly dividing the set of observations into k groups, or folds, of approximately equal size. The first fold is treated as a validation set, and the method is fit on the remaining k − 1 folds.

Properties:

1) variability is typically much lower than the variability in the test error estimates that results from the validation set approach
2) LOOCV is a special case of k-fold CV in which k is set to equal n
3) gives more accurate estimates of the test error rate than does LOOCV
4) $\mathrm{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k} \mathrm{MSE}_i.$

Classification CH4

two reasons not to perform classification using a regression method:

- a regression method cannot accommodate a qualitative response with more than two classes
- a regression method will not provide meaningful estimates of Pr(Y |X), even with just two classes.

1. *Logistic Regression:* models the probability that Y belongs to a particular category directly modeling Pr(Y = k|X = x) using the logistic function

i. Logistic Function

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

gives outputs between 0 and 1
To fit the model, we use a method called maximum likelihood

ii. Odds/ Log odds

$$\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}$$

take on any value between 0 and ∞

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$$

increasing X by one unit changes the log odds by β_1

2. *Estimating the Regression Coefficients*

Maximum Likelihood

- Likelihood Function

$$\ell(\beta_0,\beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1-p(x_{i'}))$$

The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are chosen to maximize this likelihood function.

3. *Multiple Linear Regression*

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}} \qquad \log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

4. *Multinomial Logistic Regression (more than 2 classes)*

- Select a single class to serve as the baseline

$$\Pr(Y=k|X=x) = \frac{e^{\beta_{k0}+\beta_{k1}x_1+\cdots+\beta_{kp}x_p}}{1+\sum_{l=1}^{K-1} e^{\beta_{l0}+\beta_{l1}x_1+\cdots+\beta_{lp}x_p}} \quad \text{(for k =1,….K-1)} \qquad \Pr(Y=K|X=x) = \frac{1}{1+\sum_{l=1}^{K-1} e^{\beta_{l0}+\beta_{l1}x_1+\cdots+\beta_{lp}x_p}}$$

$$\log\left(\frac{\Pr(Y=k|X=x)}{\Pr(Y=K|X=x)}\right) = \beta_{k0} + \beta_{k1}x_1 + \cdots + \beta_{kp}x_p \quad \text{(for k =1,….K-1)}$$

the log odds between any pair of classes is linear in the features

if we set epileptic seizure to be the baseline, then we can interpret βstroke0 as the log odds of stroke versus epileptic seizure, given that $x_1 = \ldots = x_p = 0$.

- Softmax, treat all K classes symmetrically

$$\Pr(Y=k|X=x) = \frac{e^{\beta_{k0}+\beta_{k1}x_1+\cdots+\beta_{kp}x_p}}{\sum_{l=1}^{K} e^{\beta_{l0}+\beta_{l1}x_1+\cdots+\beta_{lp}x_p}}$$

$$\log\left(\frac{\Pr(Y=k|X=x)}{\Pr(Y=k'|X=x)}\right) = (\beta_{k0}-\beta_{k'0}) + (\beta_{k1}-\beta_{k'1})x_1 + \cdots + (\beta_{kp}-\beta_{k'p})x_p.$$

5. *Generative Models for Classification:* model the distribution of the predictors X separately in each of the response classes, then use Bayes' theorem to flip these around into estimates for Pr(Y = k|X = x).

- **Why do we need this method?**
  - When there is substantial separation between the two classes, the parameter estimates for the LR are surprisingly unstable.
  - If the distribution of the predictors X is approximately ***normal*** in each of the classes and the ***sample size is small***, then the approaches in this section may be more accurate than logistic regression.
- **Bayes' theorem**
  - posterior probability: the probability that the observation belongs to the kth class, given the predictor value for that obs.

$$p_k(x) = \Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

  - the Bayes classifier will yield the smallest possible total number of misclassified observations, regardless of the class from which the errors stem, so we can modify the threshold.
- **Find a way to estimate $f_k(x)$**

i. Linear Discriminant Analysis for p = 1(Assumption: ***$f_k(x)$ is normal or Gaussian / $\sigma_1^2 = \cdots = \sigma_K^2$***)

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)} \qquad \delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

$$\hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2$$

  - The Bayes classifier involves assigning an observation X = x to the class for which $p_k(x)$ or the term below is largest.

$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

  - **Bayes decision boundary:** the point for which $\delta 1(x) = \delta 2(x)$

ii. Linear Discriminant Analysis for p > 1(Assumption: ***each individual predictor follows a one-dimensional normal distribution / common covariance matrix***)

$$f(x) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \mathbf{\Sigma}^{-1}(x-\mu)\right) \qquad \delta_k(x) = x^T \mathbf{\Sigma}^{-1}\mu_k - \frac{1}{2}\mu_k^T \mathbf{\Sigma}^{-1}\mu_k + \log \pi_k$$

  - Once we have estimates $\hat{\delta}k(x)$, we can turn these into estimates for class probabilities

$$\widehat{\Pr}(Y = k|X = x) = \frac{e^{\hat{\delta}_k(x)}}{\sum_{l=1}^{K} e^{\hat{\delta}_l(x)}}$$

		True class		
		− or Null	+ or Non-null	Total
Predicted class	− or Null	True Neg. (TN)	False Neg. (FN)	N*
	+ or Non-null	False Pos. (FP)	True Pos. (TP)	P*
	Total	N	P	

Name	Definition	Synonyms
False Pos. rate	FP/N	Type I error, 1−Specificity
True Pos. rate	TP/P	1−Type II error, power, sensitivity, recall
Pos. Pred. value	TP/P*	Precision, 1−false discovery proportion
Neg. Pred. value	TN/N*	

  - F1 Score: seeks a balance between Precision and Recall

$$F = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \qquad F_\beta = \frac{\beta^2 + 1}{\frac{\beta^2}{recall} + \frac{1}{precision}}$$

iii. Quadratic Discriminant Analysis(Assumption: ***normal or Gaussian / each class has its own covariance matrix***)

- LDA tends to be a better bet than QDA if there are relatively ***few training observations*** and so reducing variance is crucial.
- QDA is recommended if the ***training set is very large***, so that the variance of the classifier is not a major concern, or if the assumption of a common covariance matrix for the K classes is clearly untenable.
- LDA is a much less flexible classifier than QDA, and so has substantially lower variance. This can potentially lead to improved prediction performance.

iv. Naive Bayes(Assumption: Within the k_{th} class, the p predictors are independent.)

$$\Pr(Y = k|X = x) = \frac{\pi_k \times f_{k1}(x_1) \times f_{k2}(x_2) \times \cdots \times f_{kp}(x_p)}{\sum_{l=1}^{K} \pi_l \times f_{l1}(x_1) \times f_{l2}(x_2) \times \cdots \times f_{lp}(x_p)}$$

Linear Model Selection

1. Subset Selection: Use least squares to fit a linear model that contains a subset of predictors.

i. Best Subset Selection: (try 2^p possibilities)

- Fit all Cpk models -> Pick the best(smallest RSS/ highest R^2) among Cpk models.
- Select the best using cross validation prediction error (AIC/BIC/R^2..)

ii. Stepwise Selection —> not guaranteed to find the best possible model

a. Forward: (begin with null model, add predictors one at a time, try 1+p(p+1)/2 models)

- From k=0,…p-1:

  Consider all p-k models that augment the predictors in Mk with one additional predictors

  Choose the best(smallest RSS/ highest R^2) among these p-k models, and call it Mk+1
- Select the best model among M0,….MP, using cv prediction error(AIC/BIC/R^2..)

★ if n<p, it is possible to construct submodes M0,….Mn-1 only

b. Backward: (begin with full model, removes the least useful predictor one at a time)

- From k=p, p-1, ….1:

  Consider all k models that contain k-1 predictors

  Choose the best(smallest RSS/ highest R^2) among these p-k models, and call it Mk+1
- Select the best model among M0,….MP, using cv prediction error(AIC/BIC/R^2..)

★ Requires n > p(so the full model can be fit)

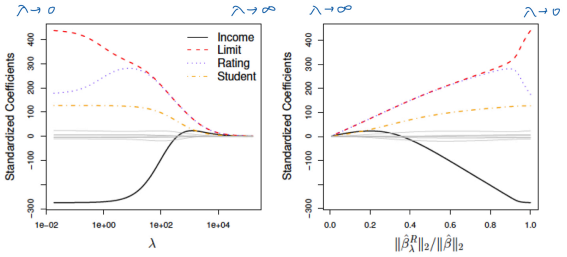
2. Shrinkage: Fit a model containing all p predictors that constrains or regularizes the coefficient estimates.

i. Ridge Regression

- the ridge regression coefficient estimates βˆR are the values that minimize

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

$$\underset{\beta}{\text{minimize}}\left\{\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2\right\}\quad\text{subject to}\quad\sum_{j=1}^{p}|\beta_j|\le s$$



- When λ = 0, the penalty term has no effect, and ridge regression will produce the least squares estimates.
- λ → ∞, the impact of the shrinkage penalty grows, and the ridge regression coefficient estimates will approach zero.
- the shrinkage penalty is applied to β1,...,βp, but not to the intercept β0.

★ Scale Equivariant: Ridge(X) <—> Least Squares(O) —> so we should apply ridge regression after standardizing the predictors

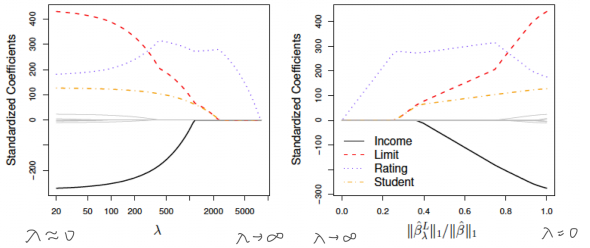
- Why Does Ridge Regression Improve Over Least Squares? -> Bias-Variance trade-off
- **Ridge Regression works best in situations where the least squares estimates have high variance.**

ii. Lasso Regression

- The lasso coefficients, βˆλL, minimize the quantity

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

$$\underset{\beta}{\text{minimize}}\left\{\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2\right\}\quad\text{subject to}\quad\sum_{j=1}^{p}\beta_j^2\le s$$



- λ → ∞, the coefficient estimates to be exactly equal to zero, yielding "sparse" models
- depending on the value of λ, the lasso can produce a model involving any number of variables.
- Can eliminating irrelevant variables, but can not handle redundant variables

one might expect :
Lasso performs better -> relatively small number of predictors have substantial coefficients, and the remaining predictors have coefficients that are very small or that equal zero.
Ridge regression performs better -> the response is a function of many predictors, all with coefficients of roughly equal size.
However, the number of predictors that is related to the response is never known a priori for real data sets.

iii. Elastic Net: Controls for correlations! Combine L1 and L2 regularizations. When α=1->L1; when α=0, L2

$$\lambda\left[\tfrac{1}{2}(1-\alpha)\|\beta\|_2^2+\alpha\|\beta\|_1\right]$$

3. *Dimension Reduction:* transform the predictors and then fit a least squares model using the transformed variables

- Let Z_1, Z_2, ..., Z_M represent M < p linear combinations of our original p predictors, for some constants φ1m,φ2m …,φpm, we can then fit the linear regression model using ordinary least squares.

$$Z_m = \sum_{j=1}^{p} \phi_{mj} X_j \qquad y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \epsilon_i, \quad i = 1, \ldots, n.$$

- Reduces the problem of estimating the p+1 coefficients to M+1 coefficients

$$\sum_{m=1}^{M} \theta_m z_{im} = \sum_{m=1}^{M} \theta_m \sum_{j=1}^{p} \phi_{jm} x_{ij} = \sum_{j=1}^{p} \sum_{m=1}^{M} \theta_m \phi_{jm} x_{ij} = \sum_{j=1}^{p} \beta_j x_{ij}, \qquad \beta_j = \sum_{m=1}^{M} \theta_m \phi_{jm}.$$

- The choice of Z_1, Z_2, ..., Z_M , or equivalently, the selection of the φ_{jm}'s, can be achieved in different ways.

i. Principal components analysis (PCA)

a. First Principal Component: linear combination of the variables with the largest variance

- The first principal component direction of the data is that along which the observations vary the most.

	X_1	X_2	....	X_p	Z_1
data 1	x_{11}	x_{12}	...	x_{1p}	z_{11}
					z_{21}
					⋮
data n	x_{n1}	x_{n2}	....	x_{np}	z_{n1}

sample variance

$$z_{i1} = \sum_{j=1}^{p} \phi_{1j} x_{ij}$$

Select ϕ_{ij}'s so that the sample variance of Z is maximized, subject to $\sum_{j=1}^{p} \phi_{ij}^2 = 1$

- The values of $z_{11}, \ldots, z_{n1}$ are known as the principal component scores.
- The first principal component vector defines the line that is as close as possible to the data.

(RSS and R2 are not suitable for selecting the best model among a collection of models with different numbers of predictors.)

Choosing the optimal model

1. Adjustment to the training error

1) Mallow's Cp: adds a penalty to training RSS to adjust that training error tends to underestimate test error (smaller better)

$$C_p = \frac{1}{n}\left(\mathrm{RSS} + 2d\hat{\sigma}^2\right) \qquad \hat{\sigma} = \mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

d is the total number of parameters used

2) AIC: If linear model with Gaussian errors, maximum likelihood and least squares are the same thing, Cp=AIC

$$\mathrm{AIC} = -2\log L + 2 \cdot d$$

L is the maximized value of the likelihood function

3) BIC: replaces the 2dσ^2 used by Cp with a log(n)dσ^2

$$\mathrm{BIC} = \frac{1}{n}\left(\mathrm{RSS} + \log(n) d\hat{\sigma}^2\right)$$

n is the number of examples, places a heavier penalty on models with many variables, and hence results in the selection of smaller models than Cp.

4) Adjust R^2: Unlike the R^2, the adjusted R2 statistic pays a price for the inclusion of unnecessary variables in the model (bigger better)

$$\text{Adjusted } R^2 = 1 - \frac{\mathrm{RSS}/(n-d-1)}{\mathrm{TSS}/(n-1)}$$

While RSS always decreases as the number of variables in the model increases, RSS/(n−d−1) may increase or decrease, due to the d

2. Directly estimate the test error: Using validation set and cross-validation method. Does not require an estimate of the error variance σ^2.

Exercise:

- ***Flexibility:*** (flexible statistical learning method to be better or worse than an inflexible method)
  - ➡ The sample size n is extremely large, and the number of predictors p is small. (B)
  - ➡ The number of predictors p is extremely large, and the number of observations n is small. (W)
  - ➡ The relationship between the predictors and response is highly non-linear. (B)
  - ➡ The variance of the error terms, i.e. σ2 = Var(ε), is extremely high. (W
- ***Parametric/ Non-Parametric***
  - Parametric: reduces the problem of estimating f down to one of estimating a set of parameters because it assumes a form for f.
  - Non-Parametric: does not assume a functional form for f and so requires a very large number of observations.
  - The advantages of a parametric approach to regression or classification are the simplifying of modeling f to a few parameters and not as many observations are required compared to a non-parametric approach.
  - The disadvantages of a parametric approach to regression or classification are a potential to inaccurately estimate f if the form of f assumed is wrong or to overfit the observations if more flexible models are used.
- ***Lasso Regression***

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p}|\beta_j|\le s$$

  - As we increase s from 0
    - *training RSS* -> Steadily decrease
    - *test RSS* -> Decrease initially, and then eventually start increasing in a U shape.
    - *variance* -> Steadily increase
    - *(squared bias)* -> Steadily decrease
    - *irreducible error* -> Remain constant
- ***Ridge Regression***

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

  - As we increase λ from 0
    - *training RSS* -> Steadily increase
    - *test RSS* -> Decrease initially, and then eventually start increasing in a U shape.
    - *variance* -> Steadily decrease
    - *(squared bias)* -> Steadily increase
    - *irreducible error* -> Remain constant
- ***Lasso/ Ridge/ Least Squares***
  - Lasso, and Ridge Regression relative to least squares, are : *Less flexible and hence will give improved prediction accuracy when its increase in bias is less than its decrease in variance.*
  - Non-linear methods relative to least squares, is : *More flexible and hence will give improved prediction accuracy when its increase in variance is less than its decrease in bias.*
- ***LDA/ QDA***
  - If the Bayes decision boundary is linear, do we expect LDA or QDA to perform better on the training set? On the test set?

    ➡If the Bayes decision boundary is linear, we expect QDA to perform better on the training set because it's higher flexiblity will yield a closer fit. On the test set, we expect LDA to perform better than QDA because QDA could overfit the linearity of the Bayes decision boundary.
  - If the Bayes decision boundary is non-linear, do we expect LDA or QDA to perform better on the training set? On the test set?

    ➡If the Bayes decision bounary is non-linear, we expect QDA to perform better both on the training and test sets.
  - In general, as the sample size n increases, do we expect the test prediction accuracy of QDA relative to LDA to improve, decline, or be unchanged? Why?

    ➡We expect the test prediction accuracy of QDA relative to LDA to improve, in general, as the the sample size n increases because a more flexibile method will yield a better fit as more samples can be fit and variance is offset by the larger sample sizes.
  - True or False: Even if the Bayes decision boundary for a given problem is linear, we will probably achieve a superior test error rate using QDA rather than LDA because QDA is flexible enough to model a linear decision boundary. Justify your answer.

    ➡False. With fewer sample points, the variance from using a more flexible method, such as QDA, would lead to overfit, yielding a higher test rate than LDA