Section 11 Estimate Prediction Errors

Cross Validations

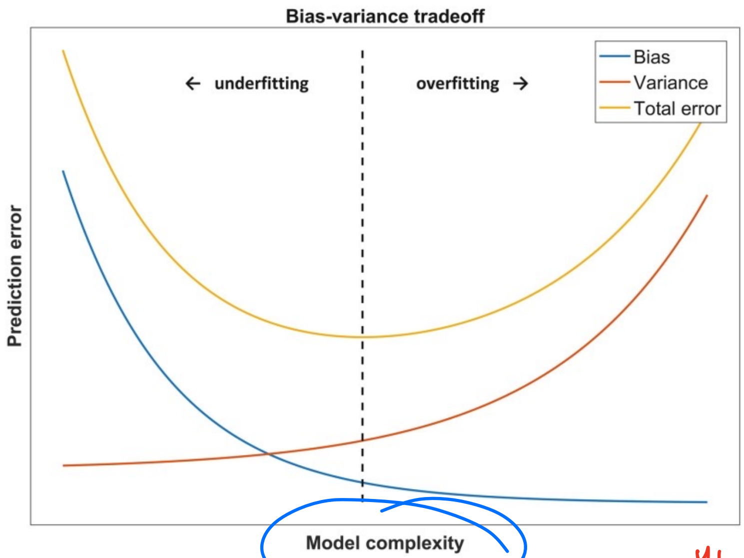
2.1 Cross validation

2.2 Leave-One-Out Cross validation

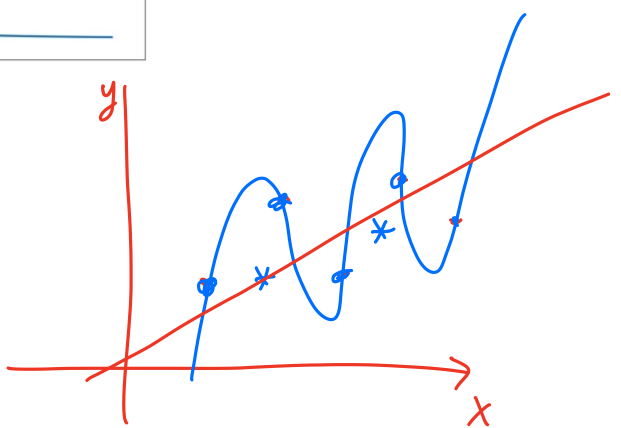
2.3 K-Cross validation

Adjusted Training Errors

- U-shaped bias–variance trade-off curve (Geman et al., 1992).



X, $\bar{y}$



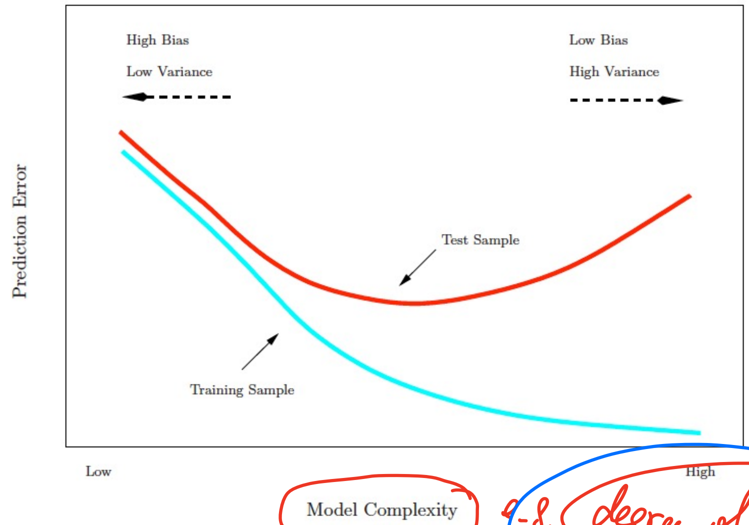
$h(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \cdots + \theta_5 x^5$

Test error V.S. Training error

The **training error** can be easily calculated by applying the statistical learning method to the observations used in its training.

The **test error** is the average error that results from using a statistical learning method to predict the response on a new observation, one that was not used in training the method.

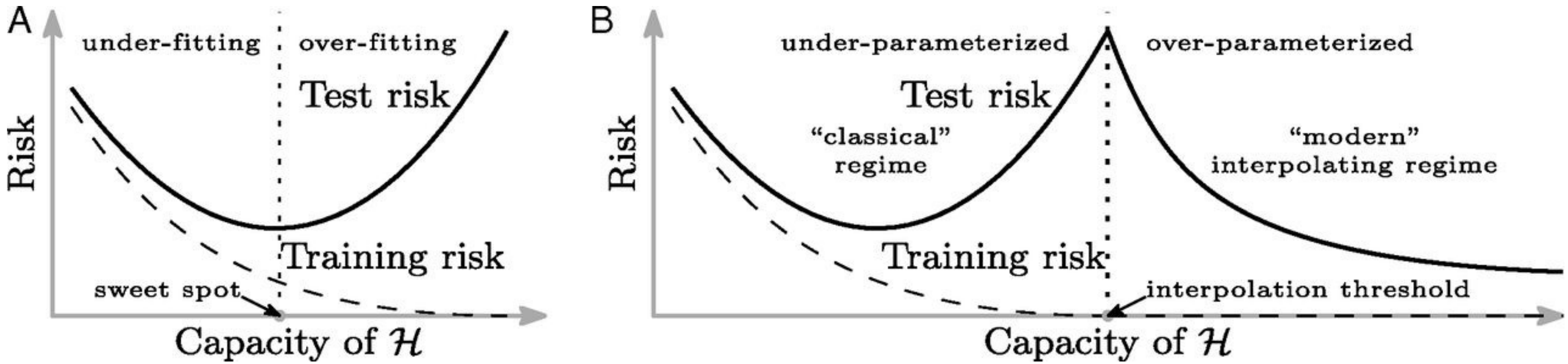
But the training error rate often is quite different from the test error rate, and in particular the former can dramatically underestimate the latter.



e.g. $J_{\text{ridge}}(\vec{\theta}) = \|X\vec{\theta} - \vec{y}\|^2 + \lambda\|\vec{\theta}\|^2$

e.g. degree of polynomial.

➢ Modern point of view of bias-variance trade-off: (Optional)



1. Reconciling modern machine-learning practice and the classical bias–variance trade-off
Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal
PNAS August 6, 2019 116 (32) 15849-15854; https://doi.org/10.1073/pnas.1903070116

2. Rethinking Bias-Variance Trade-off for Generalization of Neural Networks
Zitong Yang, Yaodong Yu, Chong You, Jacob Steinhardt, Yi Ma
Proceedings of the 37 th International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.
https://arxiv.org/pdf/2002.11328.pdf

3. **A Modern Take on the Bias-Variance Tradeoff in Neural Networks**
Neal, Mittal, Baratin, et.al. https://arxiv.org/pdf/1810.08591.pdf

Prediction-error estimates

Our ultimate goal is to produce the best model with best prediction accuracy.

1. We consider a class of validation methods that estimate the test error, by holding out a subset of the training observations from the fitting process, and then applying the statistical learning method to those held out observations. The resulting validation-set error provides an estimate of the test error.

2. Some methods (adjusted R^2, the C_p statistic, AIC and BIC) make a mathematical adjustment to the training error rate in order to estimate the test error rate.

Cross validation

Training error is easily computable with training data. However, the possibility of overfit makes it cannot be used to properly assess test error.

When we have enough data, we can *randomly* split the data into three parts:

Test.

- **Training data:** used to train various models.
- **Validation data:** used to assess the models and identify the best.
- **Test data:** test the results of the best model. (Optional)

paper

Data

FDA

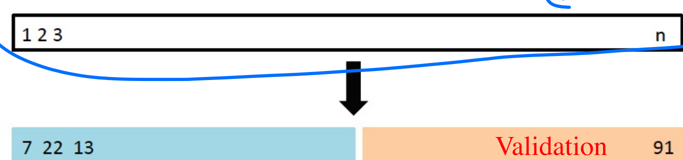
Fit various regression models on the training sample. The validation set error rates result from evaluating their performance on the validation sample. MSE as a measure of validation set error.

$X \quad \vec{y}$

$(X_{train} \quad \vec{y}_{train})$

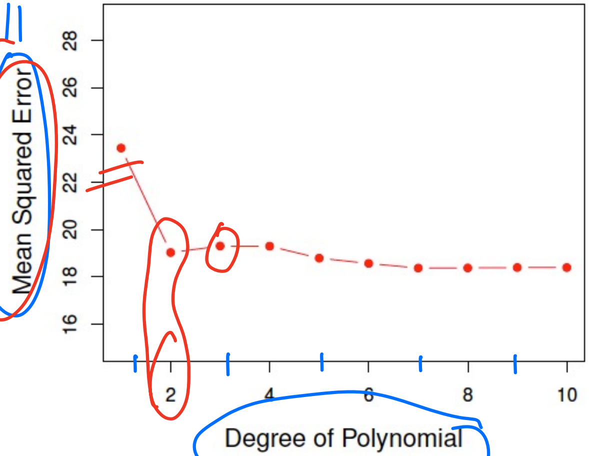
$(X_{test} \quad \vec{y}_{test})$

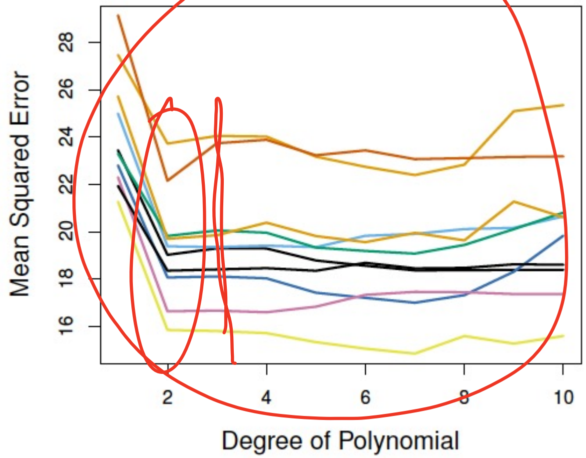
randomly



$\vec{\theta} = (X_{tr}^T X_{tr})^{-1} X_{tr}^T \vec{y}_{tr}$

$\frac{1}{n}RSS$





$$\frac{1}{n}RSS(\vec{\theta}) = \frac{1}{n}\left(X_{test}\vec{\theta}-\vec{y}_{test}\right)^T\left(X_{test}\vec{\theta}-\vec{y}_{test}\right)$$

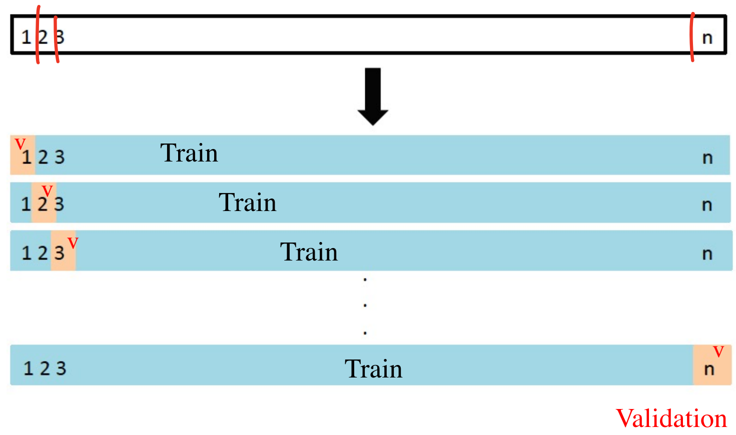
Left: Validation error estimates for a single split into training and validation data sets.

Right: The validation method was repeated ten times, each time using a different random split of the observations into a training set and a validation set. This illustrates the variability in the estimated test MSE that results from this approach

The leave-one-out cross-validation (LOOCV)

small data set n=100.

First, pick data point 1 as validation set, the rest as training set. Fit the model on the training set, evaluate the test error, on the validation set, denoted as MSE_1.



Second, pick data point 2 as validation set, the rest as training set. Fit the model on the training set, evaluate the test error on the validation set, denoted as say MSE_2.

…

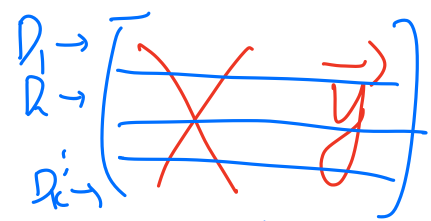
Repeat the procedure for all data point.

…

Obtain an estimate of the test error by combining the MSE_i for $i = 1,2, \ldots n$.

k=n

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} MSE_i$$

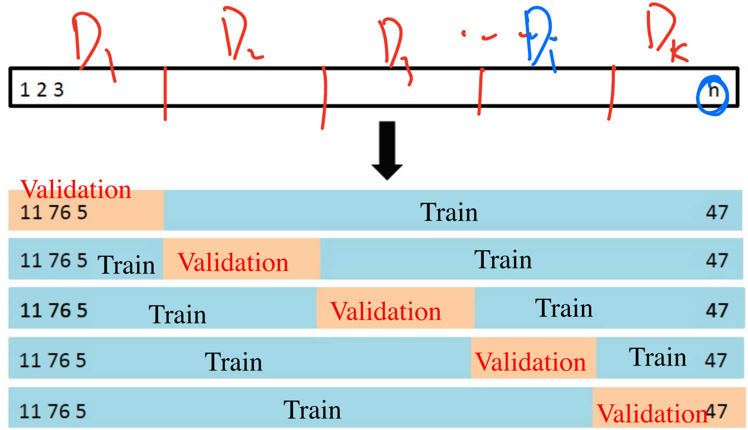


$h(\vec{x})$ $(X_i \;\; \vec{y}_i)$

- **K-fold cross validation (**widely used approach for estimating test error**)**

Divide the data (randomly) into K subsets, usually of equal or similar sizes $\frac{n}{K}$.

Data



Treat one subset as validation set, the rest together as a training set. Run the model fitting on training set. Calculate the test error estimate on the validation set, denoted as MSE_i.

...

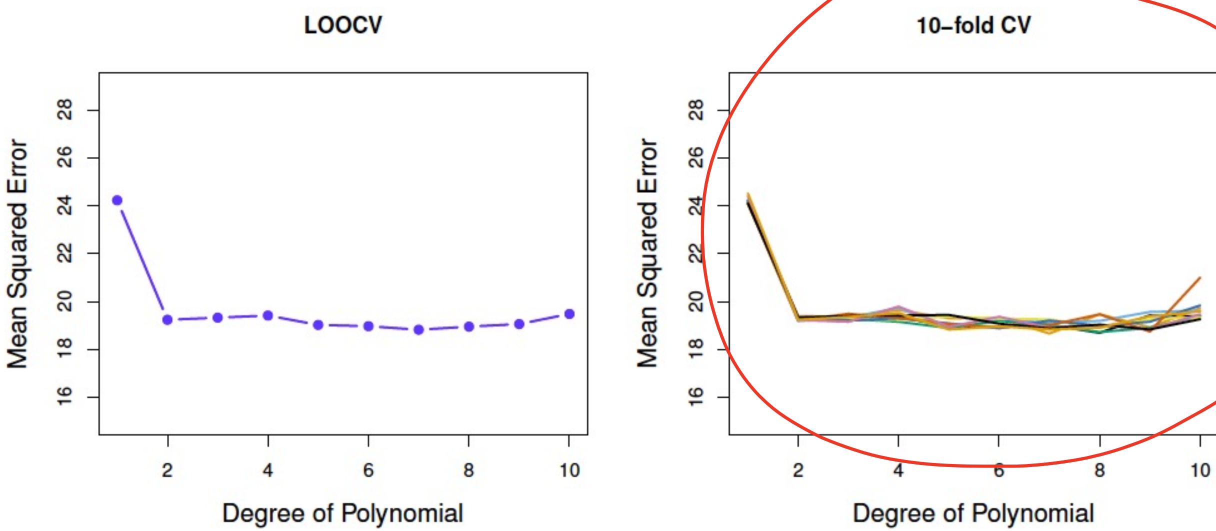
$$MSE_i = \frac{1}{n/K} \left\| h(X_i) - \vec{y}_i \right\|^2$$

Repeat the procedures over every subset.

...

Average over the above K estimates of the test errors, and obtain

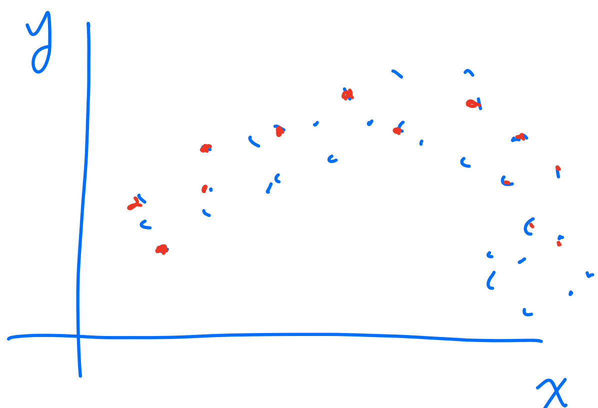
$$CV_{(K)} = \frac{1}{K} \sum_{i=1}^{K} MSE_i$$



Left: The LOOCV error curve.

Right: 10-fold CV was run nine separate times, each with a different random split of the data into ten parts.

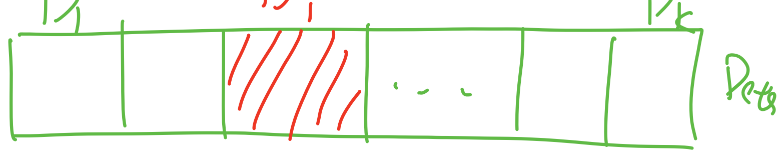
Ex:



try $f(x) = \theta_0 + \theta_1 x$

try $f(x) = \theta_0 + \theta_1 x + \theta_2 x^2$

$\vdots$



➢ Cross validation for classification

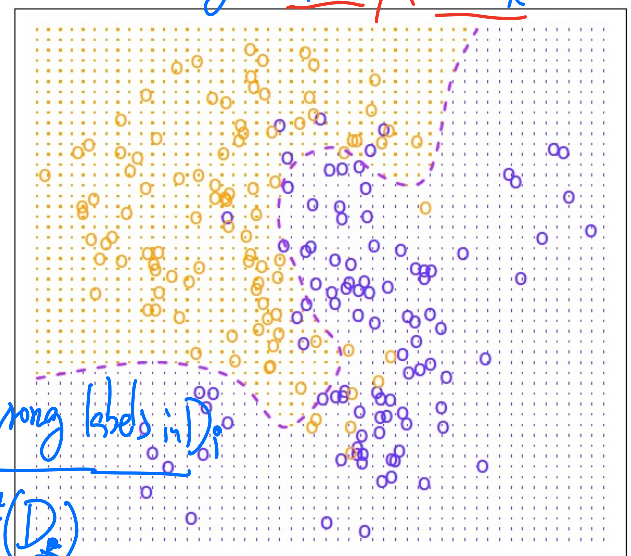
For classification with qualitative response, a natural choice is: 1 for incorrect classification and 0 for correct classification.

$f_i(\ )$ is the model by $D_1 \cdots \cancel{D_i} \cdots D_k$

We divide the data (randomly) into K equal-sized subsets.

$$CV_{(K)} = \frac{1}{K}\sum_{i=1}^{K} Mean\ Error_{(i)}$$

$$Mean\ Error_{(i)} := \frac{\sum_{(\vec{x},y)\in D_i} \mathbb{1}(f_i(\vec{x})\neq y)}{n/k} = \frac{\text{wrong labels in } D_i}{\#(D_i)}$$



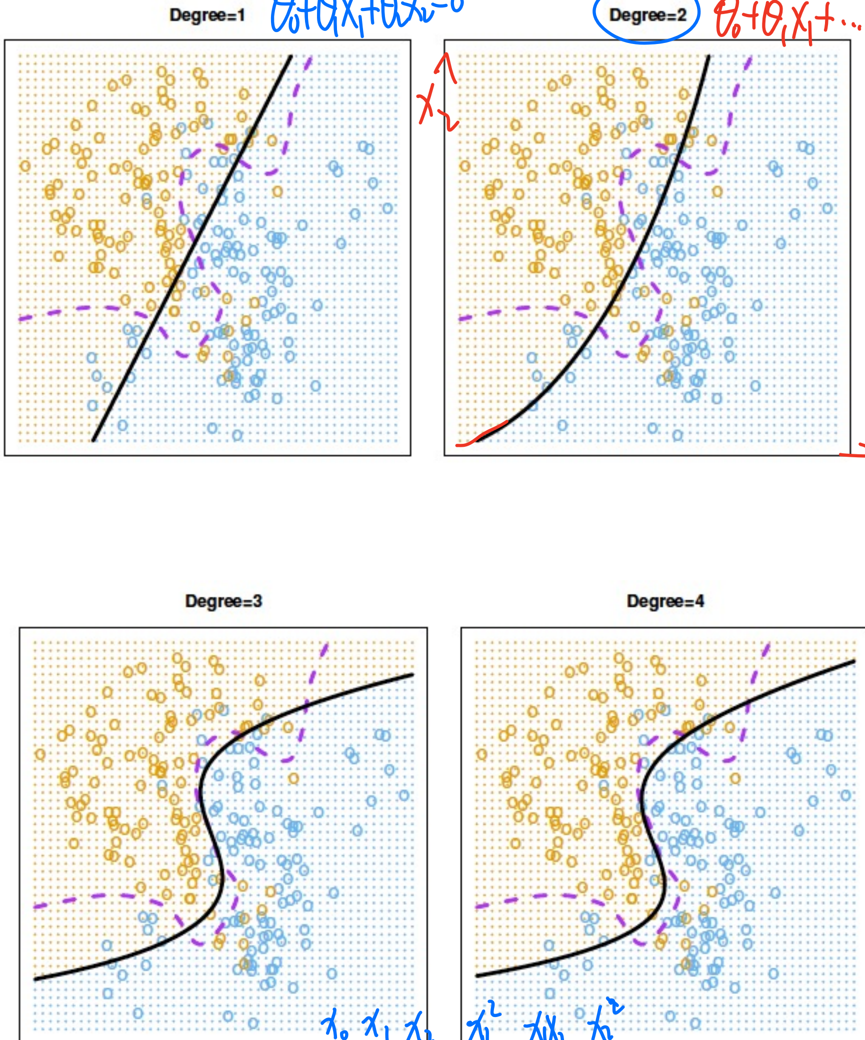
A simulated data set consisting of 100 observations in each of two groups, indicated in blue and in orange. The purple dashed line represents the Bayes decision boundary. The orange background grid indicates the region in which a test observation will be assigned to the orange class, and the blue background grid indicates the region in which a test observation will be assigned to the blue class.

$$h(\vec{x}) := \frac{1}{1+e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2)}}$$

$1+e^{(\theta_0+\theta_1 x_1+\theta_2 x_2+\theta_3 x_1^2+\theta_4 x_1 x_2+\theta_5 x_2^2)}$

$\theta_0+\theta_1 x_1+\theta_2 x_2=0$

$\theta_0+\theta_1 x_1+\cdots+\theta_5 x_2^2=0$



The Bayes decision boundary is represented using a purple dashed line.

Estimated decision boundaries from linear, quadratic, cubic and quartic (degrees 1 to 4) logistic regressions are displayed in black.

The (true) test error rates for the four logistic regression fits are respectively 0.201, 0.197, 0.160, and 0.162, while the Bayes error rate is 0.133.

In practice the true population distribution is unknown. Thus, the true test error cannot be computed. We use cross validation to solve the problem.

$\theta_0+\theta_1 z_1+\cdots+\theta_5 z_5=0$

$z_1=x_1$

$z_2=x_2$

$z_3=x_1^2$

$z_4=x_1x_2$

$z_5=x_2^2$

Data

$$X=\begin{array}{cccccc} x_0 & x_1 & x_2 & x_1^2 & x_1x_2 & x_2^2 \\ \hline 1 & \bullet & \ast & \bullet^2 & \bullet\ast & \ast^2 \\ 1 & \bullet & \ast & \bullet^2 & \bullet\ast & \ast^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{array} \qquad \vec{y}=\begin{bmatrix}1\\0\\1\\0\\\vdots\end{bmatrix}$$

More error prediction methods.

Mean Square Error $MSE = \frac{RSS}{n}$

$\boxed{SS_{total} \text{ is for } \boxed{y = \theta_0} = \bar{y}}$

0. R-squared

$\boxed{RSS}$

bad

There is human readable scoring statistic is **R-squared** calculated by

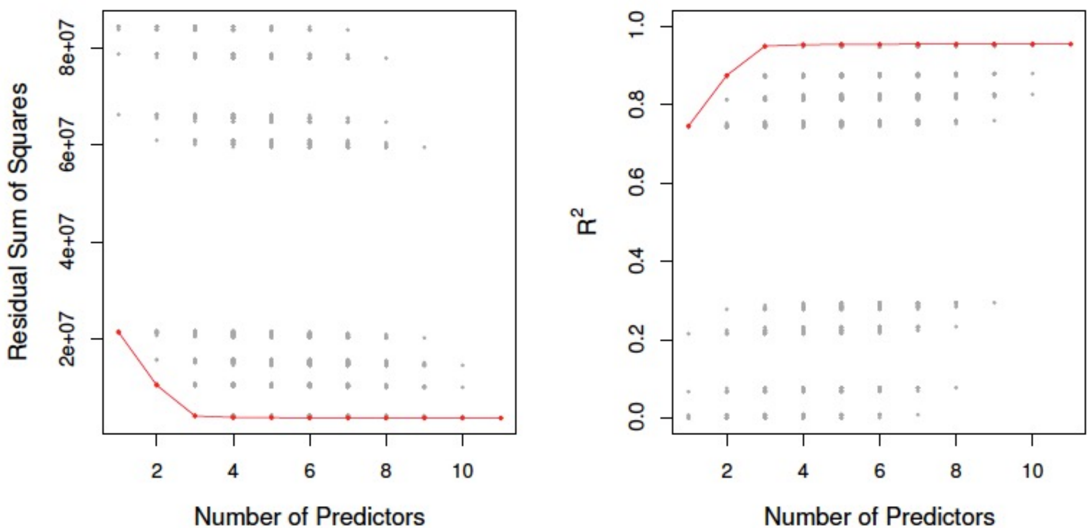
$$0 < R^2 = 1 - \frac{RSS}{SS_{total}} = 1 - \frac{RSS}{\sum_{i=1}^{n}(y^{(i)} - \bar{y})^2} < 1$$

RSS for model $f(\vec{x})$

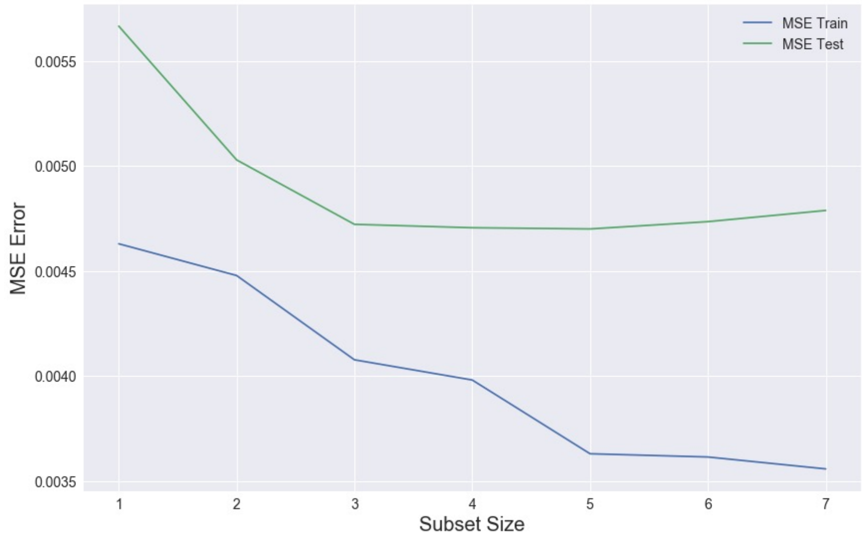
$\theta_0 + \theta_1 x_1 + \cdots \theta_d x_d$

good

So $R^2 = 1$ is perfect correlation.



The MSE and R-squared reflects the training error. However, a model with larger R-squared/ or smaller MSE error is not necessarily better than another model with smaller R-squared when we consider test error!



Modify RSS or R^2

1. Adjusted R-squared. Suppose we check k features

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

$$\textbf{adjusted } R^2 := 1 - \frac{RSS/(n-k-1)}{\sum_{i=1}^{n}(y^{(i)} - \bar{y})^2 / (n-1)}$$

With more inputs, the R^2 always increase, but the adjusted R^2 could decrease since more irrelevant inputs are penalized. The adjusted R-squared is preferred over the R-squared in evaluating models.

d features $x_1 \cdots x_d$

2. Mallows' C_p. k

take k features

The statistic of Mallow's C_p is defined as

$$\text{Mallows' } \boldsymbol{C_p} := \frac{1}{n}\left(RSS(k) + 2ks_k^2\right)$$

Here, $s_k^2 = \frac{RSS}{n-k-1}$ and $RSS(k)$ is the RSS with k features.

The model with the **smallest** C_p is preferred.

3. Akaike information criterion (AIC) $= \frac{1}{n}\left(RSS(k) + \underline{penalty'}\right)$

4. Schwarz's Bayesian information criterion(BIC) $= \frac{1}{n}\left(RSS(k) + \underline{penalty''}\right)$

The model with the **smallest AIC or** BIC is preferred.