$Y = f(X) + \varepsilon$ ($E(\varepsilon)=0$) $X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ Properties:

random fixed random fixed

1) Prediction: point prediction & predictive distribution
2) Inference: linear or not? x and y relation? influcial predictors?

Assume 1) known ε 2) known true generating function $f(x)$

optimal $f(x) = E(Y|X=x)$ minimizes MSPE: $E((Y-g(x))^2 | X=x)$ over all $g(x)$ at all points $X=x$.

For any estimated $\hat{f}(x)$ of $f(x)$, $MSPE = \underbrace{E[f(x)-\hat{f}(x)]^2}_{\text{reducible}} + \underbrace{Var(\varepsilon)}_{\text{irreducible}}$, for $\varepsilon = Y - f(x)$

reducible error = 0 if no model misspecification when n is large

irreducible even if we know $f(x)$

K-Nearest Neighbor: $\hat{f}(x) = Ave(Y|X \in N(x))$, loss function, setting $g(x) = f(x) = E(Y(x))$

$N(x)$ is neighborhood of x when no data points at $X=x$.

if we know the data generating function

↓ perform bad for large p (# of parameters) bc curse of dimensionality with simple func.)

Bias: the distance between the predictions of a model and the true values (error from approximate model

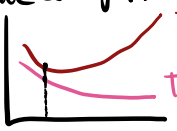
Variance: how spread the data is | how much the predictor $\hat{f}$ change when we have new data set.

Bias-Variance Trade Off: complicated & flexible model → high variance & low bias → overfitting

simple linear model → low variance & high bias → under-fitting & good interpretability

$MSE_{Tr} = Ave_{i \in Tr}[y_i - \hat{f}(x_i)]^2$

$MSE_{Te} = Ave_{i \in Te}[y_i - \hat{f}(x_i)]^2$



MSPE prefers overfitted model, typically decrease when complexity increases. bias decreases, variance increases.

if low noise ($\varepsilon \downarrow$), less dangerous in overfitting.

Bayes optimal classifer: $C(x) = j$ if $p_j(x) = \max\{p_1(x), p_2(x), \dots, p_K(x)\}$ minimizes MCE.

↓ has smallest value in missclasification error $Err_{Te} = Ave_{i \in Te} I[y_i \neq \hat{C}(x_i)]$

KNN to estimate classification: small K, overfitting | large K, underfitting

↓ controls the complexity of the classifier

$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$ $\log(P(x)/1-p(x)) = \beta_0 + \beta_1 x$

Likelihood: find β_0, β_1 parameters to maximize the probability of drawing the observed data.

training $MSPE = RSS = e_1^2 + e_2^2 + \dots + e_n^2$ for $e_i = y_i - \hat{y}_i$ t-statistics $= \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$ $n-2$ degrees of freedom ↓ estimating 2 parameters from the data.

$RSE = \sqrt{\frac{1}{n-2}RSS} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$ estimate of noise standard deviation → sqrt(Var(ε))

$R^2 = 1 - \frac{RSS}{TSS}$, % of response variation captured by fitted model, always increase with more parameters highest with model of all predictors.

Adjusted $R^2 = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}$ RSS always decrease, R^2 always increase with more predictors, because they are about training error adjusted R^2 penalize model with many predictors and decrease with unnecessary predictor

How to choose good model?

- Best-subset selection: 2^p times of computation, not generalizable, overfit with training data
- Forward Stepwise selection: only suitable model for large p computational advantages, not guarantee for best predicted model.
- Backward Stepwise selection: n must > p

Interpretation: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$ or $\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 \cdot X_2)$

① β_1: the average effect on $\hat{Y}$ of a one-unit increase in X_1 is always β_1, holding all other predictors fixed. note multi-collinearity

② The coefficient estimates in the table suggest that an increase in X_1 of XX is associated with increase $\hat{Y}$ of $(\hat{\beta}_1 + \hat{\beta}_3 \times X_2) \times XX$ = ____ units. (note hierarchy principle)

$\log\left(\frac{P(x)}{1-P(x)}\right) = \hat{\beta}_0 + \hat{\beta}_1 x$,

$P(x) = \frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)}$

① $\hat{\beta}_1$ = the change in predicted log odds associated with one unit increase in X.

② one unit increase in x on the probability: A one-unit increase in x results in an increase of β for the log-odds of P(x)

③ $\exp\{\hat{\beta}_1\}$: the multiplicative change in predicted odds associated with one unit ↑ in X.

④ $\hat{\beta}_0$ and $\exp\{\hat{\beta}_0\}$: the estimated log odds and odds when $X_i = 0$

Test Error Method: ① indirectly adjust the training error (AIC, BIC, C_p) ② directly estimate the test error (CV, validation)

$BIC = \frac{1}{n}(RSS + \log(n) d\hat{\sigma}^2)$ — penalize large p, result in simplified model
$C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$ — only for linear regression, clear criterion
$AIC = -2\log L + 2d$ — for non-linear regression, interpretable

Validation & Cross-Validation provide an direct estimation, no need for $\hat{\sigma}^2$, df

$MSPE = bias^2 + var$

validation set: fit on training set, fitted model is used to predict the responses for the observations in the validation set.

problems: ① high variance in estimating test error ③ bias from data drawing
② high variability for optimal complexity (each split gives very different conclusions)
④ overestimate the test error bc estimating a model with full data set by a model with a few points.

K-fold Cross-validation: randomly divide data into different K equal-size parts. For classification problem:

$CV_{(K)} = \sum_{k=1}^{K} \frac{n_k}{n} MSE_k$, where $MSE_k = \sum_{i \in C_k} (y_i - \hat{y}_i)^2 / n_k$ and $\hat{y}_i$ is the fit for observation i $CV_k = \sum_{k=1}^{K} \frac{n_k}{n} Err_k$

LOOCV setting $K = n$ yields n-fold, no randomness, high variance because individual test error closed related.

model fitting trade-off:
- estimate regression function $f(x)$
- assume test error curve is known
- more complex model, higher var low bias, less interpretability.

test error estimation Trade-off:
- estimate test error from data
- trade-off for k-fold CV:
large K → high variance, low bias
small K → high bias, low var

LOOCV: low bias → training data close to test data
high var → fewer testing data for E(MSPE) estimation

5-fold CV: high-bias → less training data compare with the full data set.
low var → more testing data / less correlation between models.

Decomposition of MSPE:

$E[(Y - \hat{f}(x))^2] = E[(Y - E(Y|X) + E(Y|X) - \hat{f}(x))^2] = E[(Y - E(Y|X))^2] + E[(E(Y|X) - \hat{f}(x))^2]$

$+ 2E[(Y - E(Y|X))(E(Y|X) - \hat{f}(x))] \Rightarrow E[E[Y - E[Y|X]|X](E[Y|X] - \hat{f}(x)] = 0$

$= Bias^2(x) + Var(x)$ so the MSPE minimizing solution is to choose $\hat{f}(x) = E[Y|X]$ ↙ but we don't know the distribution of (X, Y)