ECE7252
Statistical Learning for Signal Processing

Lecture 8: Linear Methods for Regression (Part I: Theory and Algorithm)

Chin-Hui Lee
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, GA 30332, USA
chl@ece.gatech.edu

Lecture Outline on Regression

- Statistical learning: general concept
- Supervised learning: learning with a teacher
- Regression and classification problems
- Model selection, feature selection and generalization
- Autoregression
- Statistical decision theory (in Lecture on classification)

An Objective of Regression

- Goal : find a "good" model $f(X)$ to predict a variable Y from a set of predictors, $X_1, X_2 \ldots X_p$
  - Quality measure: expected prediction error = $E((Y-f(X))^2)$
  - Solution: take a statistic, $f(x) = E(Y|X=x)$, but … what is $E(Y|X=x)$?
  - Decomposition of the expected prediction error EPE

If $Y = E(Y|X) + e$ with $Var(\varepsilon) = \sigma^2$ and $\hat{f}(X)$ is an estimator of the model

$$EPE = E_{Y|X=x}((y-\hat{f}(x))^2)$$

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

$$= \sigma^2 + Var(\hat{f}(x)) + \text{Biais}^2$$

Mean Square Error

Linear Regression Models

- Linear regression models suppose that *E(Y|X)* is linear

$$E(Y \mid X) = \beta_0 + \sum_{i=1}^{N} \beta_i X_i = f(X)$$

- Linear models are classical tools but …
  - Simple and effective
  - Allow an easy interpretation of regressors (predictor) effects
  - General since X_i's can be any function of other variables (linear or nonlinear, quantitative or qualitative)
  - Useful to understand because most other methods are generalisations of them

Linear Regression Formulation

- Least Squares: Minimizing Sum of Squared Error

$$D = \sum_{t=1}^{n} d_i^2 = \sum_{t=1}^{n} [y_i - (a + bx_i)]^2 = \text{minimum}$$

- We obtain the following matrix normal equation

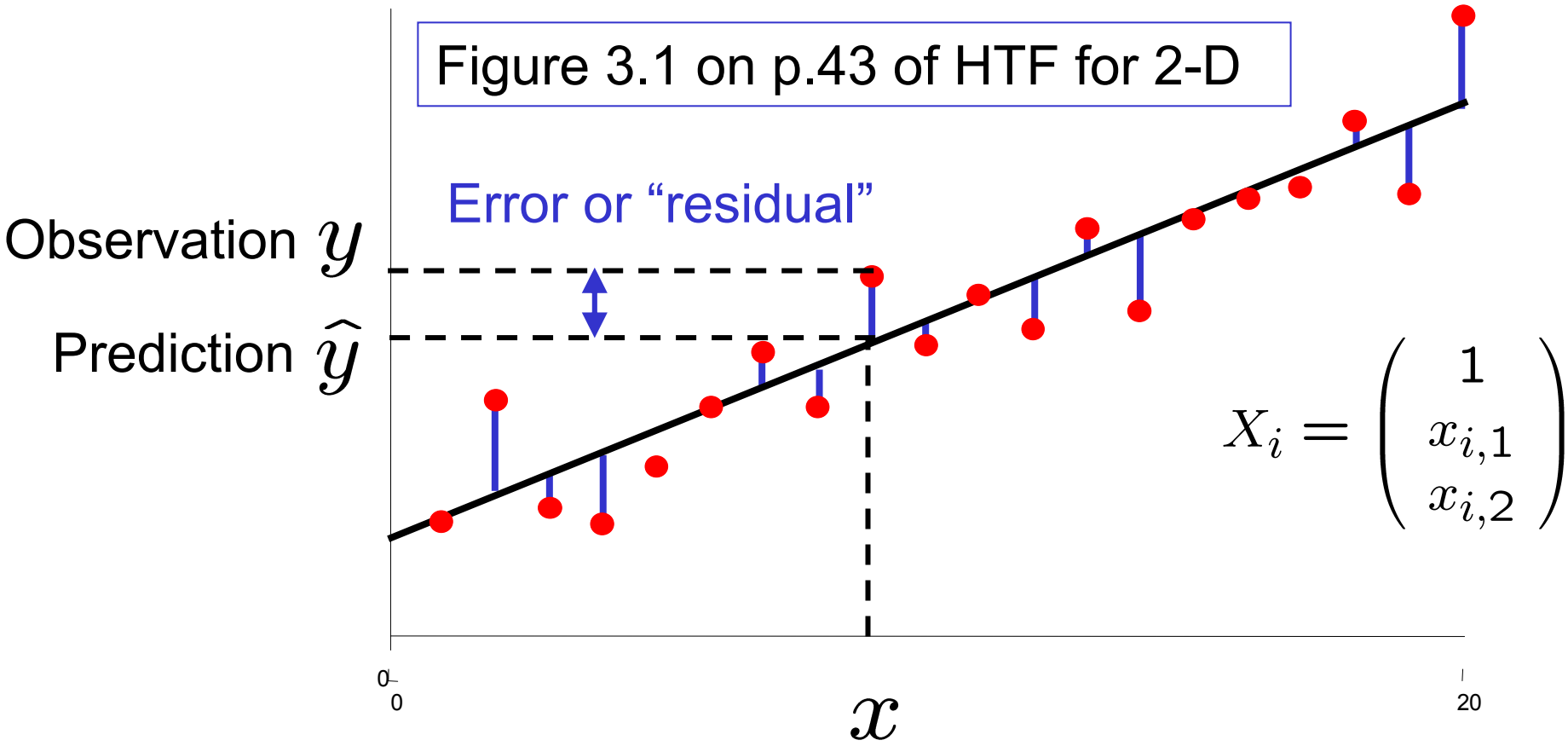
$$\frac{\partial D}{\partial a} = 0 \Rightarrow \sum_{t=1}^{n} y_i = an + b\sum_{t=1}^{n} x_i, \quad \frac{\partial D}{\partial b} = 0 \Rightarrow \sum_{t=1}^{n} x_i y_i = a\sum_{t=1}^{n} x_i + b\sum_{t=1}^{n} x_i^2$$

- Solving for intercept *a* and slope *b* : y=polyfit(y,x,n)

$$\hat{b} = \frac{n\sum_{t=1}^{n} x_i y_i - (\sum_{t=1}^{n} x_i)(\sum_{t=1}^{n} y_i)}{n\sum_{t=1}^{n} x_i^2 - (\sum_{t=1}^{n} x_i)^2}, \hat{a} = \frac{(\sum_{t=1}^{n} y_i)(\sum_{t=1}^{n} x_i^2) - (\sum_{t=1}^{n} x_i)(\sum_{t=1}^{n} x_i y_i)}{n\sum_{t=1}^{n} x_i^2 - (\sum_{t=1}^{n} x_i)^2} = \frac{\sum_{t=1}^{n} y_i - \hat{b}\sum_{t=1}^{n} x_i}{n} = \bar{\bar{Y}} - \hat{b}\bar{\bar{X}}$$

- Extend to more than one regressor (econometrics)

Least Squares Fitting



Sum of squared error: $L(w) = \sum_{i=1}^{n} (y_i - \boldsymbol{w}^\top \boldsymbol{x}_i)^2$

Least Square Estimation Solutions

- Least square is the most popular method to estimate linear models by minimizing "residual" sum of squares:

$$\arg\min_{\beta} RSS(\beta) = \arg\min_{\beta} \sum_{i=1}^{N} (y_i - f(x_i))^2$$

- Normal equation: $X'(Y - X\hat{\beta}) = 0$
- Solution: $\hat{\beta} = (X'X)^{-1}X'Y$ (if psuedoinverse $(X'X)^{-1}$ exists)
- Predicted values: $\hat{y} = X\hat{\beta} = X(X'X)^{-1}X'Y = HY$
  - The predictions are an orthogonal projections of *y* in the subspace of R^n defined by the columns of the *X* matrix. *H* is the projection matrix. (refer to your textbook on linear algebra)
- The method can also be generalized to multiple *Y*'s
  - If the set of *k variables, Y_i*'s, are independent, the multivariate solution is identical to doing *k* separate classical regressions

Multiple Regression

- Multiple regression parameters can be estimated from a sequence of simple univariate regressions with

$$\hat{\beta} = \frac{x'y}{x'x} = \frac{< x, y >}{< x, x >}$$

  - If *columns of X are orthogonal*, the parameters estimates are simply simple regressions of *Y* on the columns of *X*
  - The *algorithm of successive orthogonalization* allows to obtain the estimator of b_i in a multiple regression as the estimator of the parameter of a univariate regression of *Y* on the residual z_i of the regression of X_i on the other columns of *X*. z_i is the part of X_i which is orthogonal to the space of the other columns of *X* and is obtained by a sequence of successive orthogonalizations
  - The *Gram-Schmidt* procedure is a numerical strategy based on the previous algorithm to compute LS estimates

Properties of Least Squares Estimators

- If Y_i are independent, X fixed and $V(Y_i)=s^2$ constant:

$$E(\hat{\beta})=\beta \quad V(\hat{\beta})=\sigma^2(X'X)^{-1} \quad \text{and} \quad \mathrm{E}(\hat{\sigma}^2)=\sigma^2$$

$$\text{with } \hat{\sigma}^2=\frac{1}{N-p-1}\sum(y_i-\hat{f}(x_i))^2$$

- If, in addition, $Y=f(X_i)+e$ with $e \sim N(0, s^2)$:

$$\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1}) \quad \text{and} \quad (N\text{-}p\text{-}1)\hat{\sigma}^2 \sim \sigma^2 \chi^2_{N-p-1}$$

  - Normal assumption is common but not necessarily valid
  - T tests can be built to test the nullity of any one parameter
  - F tests can be built to test the nullity of a vector of parameters (or linear combinations of them)
  - Confidence intervals can be built for parameters

Gauss-Markov Theorem

- Gauss-Markov Theorem
  - Refer to any textbook on linear models
  - Does not need normal assumption on errors, as long as they are independent with the same variance
  - The least square estimate of the parameters has the smallest variance among all linear UNBIASED estimators
    - LS estimators do not have the small mean square error if one accepts biased estimators (e.g. ridge regression)
    - And since all model are often wrong, LSE will always be biased

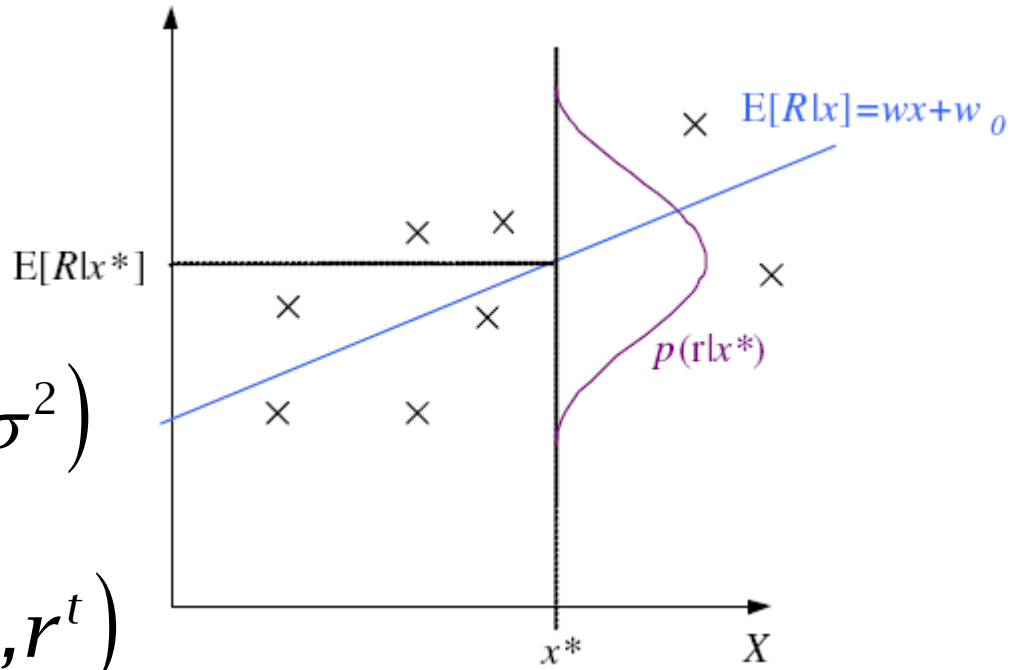
Regression with Gaussian Errors

$$r = f(x) + \varepsilon$$

$$\text{estimator}: g(x \mid \theta)$$

$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$

$$p(r \mid x) \sim \mathcal{N}(g(x \mid \theta), \sigma^2)$$



$$\mathcal{L}(\theta \mid \mathcal{X}) = \log \prod_{t=1}^{N} p(x^t, r^t)$$

$$= \log \prod_{t=1}^{N} p(r^t \mid x^t) + \log \prod_{t=1}^{N} p(x^t)$$

Regression as Maximum Likelihood

$$\mathcal{L}(\theta \mid \mathcal{X}) = \log \prod_{t=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[- \frac{\left[r^t - g\left(x^t \mid \theta\right)\right]^2}{2\sigma^2} \right]$$

$$= -N \log \sqrt{2\pi}\sigma - \frac{1}{2\sigma^2} \sum_{t=1}^{N} \left[r^t - g\left(x^t \mid \theta\right)\right]^2$$

$$E(\theta \mid \mathcal{X}) = \frac{1}{2} \sum_{t=1}^{N} \left[r^t - g\left(x^t \mid \theta\right)\right]^2$$

LS Regression and ML Solution

$$g\left(x^t \mid w_1, w_0\right) = w_1 x^t + w_0$$

$$\sum_t r^t = N w_0 + w_1 \sum_t x^t$$

$$\sum_t r^t x^t = w_0 \sum_t x^t + w_1 \sum_t \left(x^t\right)^2$$

$$\mathbf{A} = \begin{bmatrix} N & \sum_t x^t \\ \sum_t x^t & \sum_t \left(x^t\right)^2 \end{bmatrix} \quad \boldsymbol{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} \quad \boldsymbol{y} = \begin{bmatrix} \sum_t r^t \\ \sum_t r^t x^t \end{bmatrix}$$

$$\mathbf{w} = \mathbf{A}^{-1}\mathbf{y} \quad \Rightarrow \quad \text{same solution as LS}$$

Other Methods to Estimate Linear Models

- 3 reasons why we are not happy with least squares
  - *Prediction accuracy*: LS often provide predictions with low bias but high variance.
  - *Interpretation*: when the number of regressors is too high, the model is difficult to interpret. One seek to find a smaller set of regressors with higher effects
  - Non-Gaussian errors and outliers are common in real world
- Proposed approaches
  - Subset selection: best subset, forward, backward, stepwise
  - Shrinkage methods: ridge regression and Lasso method (+ generalization to a Bayes view of them)
  - Methods using derived input directions: principal component regression, partial least squares and canonical correlations

LS and Other Error Measures

- Square Error:

$$E(\theta \mid \mathcal{X}) = \frac{1}{2}\sum_{t=1}^{N}\left[r^t - g\left(x^t \mid \theta\right)\right]^2$$

- Relative Square Error:

$$E\left(\theta \mid \mathcal{X}\right) = \frac{\sum_{t=1}^{N}\left[r^t - g\left(x^t \mid \theta\right)\right]^2}{\sum_{t=1}^{N}\left[r^t - \bar{r}\right]^2}$$

- Absolute Error: $E\left(\theta|\mathcal{X}\right) = \sum_t |r^t - g(x^t|\theta)|$
- ε-sensitive Error:

$$E\left(\theta|\mathcal{X}\right) = \sum_t 1(|r^t - g(x^t|\theta)|>\varepsilon)\ (|r^t - g(x^t|\theta)| - \varepsilon)$$

Bias and Variance

$$E\left[(r-g(x))^2 \mid x\right] = E\left[(r-E[r \mid x])^2 \mid x\right] + (E[r \mid x]-g(x))^2$$

noise *squared error*

$$E_X\left[(E[r \mid x]-g(x))^2 \mid x\right] = (E[r \mid x]-E_X[g(x)])^2 + E_X\left[(g(x)-E_X[g(x)])^2\right]$$

bias *variance*

Estimating Bias and Variance

M samples $\mathcal{X}_i=\{x^t_i, r^t_i\}, i=1,...,M$

are used to fit $g_i(x), i=1,...,M$

$$\text{Bias}^2(g) = \frac{1}{N}\sum_t \left[\bar{g}(x^t) - f(x^t)\right]^2$$

$$\text{Variance}(g) = \frac{1}{NM}\sum_t\sum_i \left[g_i(x^t) - \bar{g}(x^t)\right]^2$$

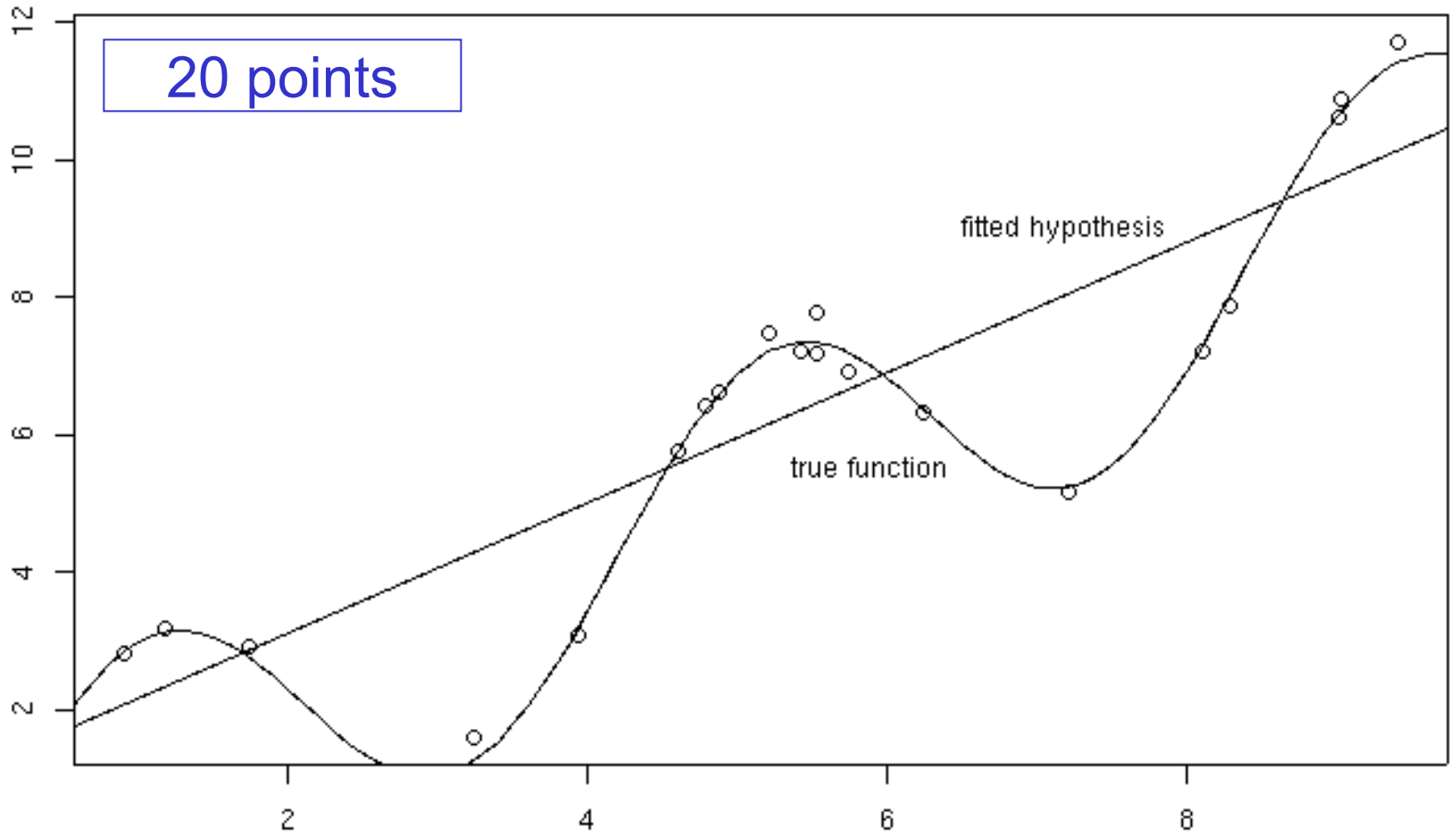
$$\bar{g}(x) = \frac{1}{M}\sum_t g_i(x)$$

Bias/Variance Dilemma

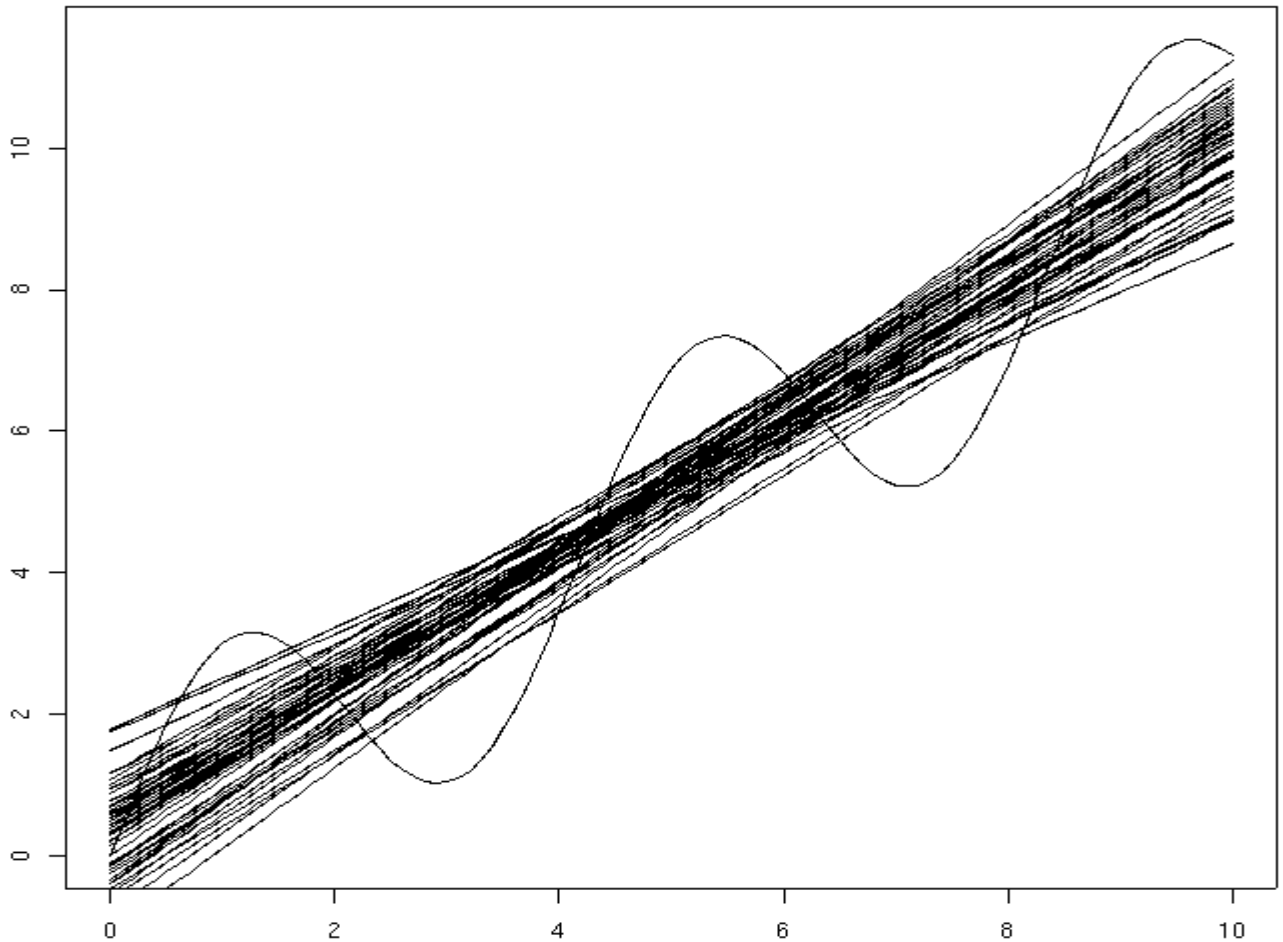
- Example: $g_i(x)=2$ has no variance and high bias
  $g_i(x)= \sum_t r^t_i /N$ has lower bias with variance

- As we increase complexity
  - Bias decreases (a better fit to data) and
  - Variance increases (fit varies more with data)
- Bias/Variance dilemma: (Geman et al., 1992)

Example: $y = x + 2\sin(1.5x) + N(0,0.2)$



Linear Regression: 50 Fits



Classical Statistical Analysis

- Imagine that our particular training sample S is drawn from some population of possible training samples according to P(S).
- Compute: $E_P [(y^* – h(x^*))^2]$
- Decompose this into “bias”, “variance”, and “noise”

Lemma

- Let Z be a random variable with probability distribution $P(Z)$
- Let $\underline{Z} = E_P[\, Z \,]$ be the average value of Z
- Lemma: $E[\, (Z - \underline{Z})^2 \,] = E[Z^2] - \underline{Z}^2$

$$E[\, (Z - \underline{Z})^2 \,] = E[\, Z^2 - 2\, Z\, \underline{Z} + \underline{Z}^2 \,]$$
$$= E[Z^2] - 2\, E[Z]\, \underline{Z} + \underline{Z}$$
$$= E[Z^2] - 2\, \underline{Z}^2 + \underline{Z}^2$$
$$= E[Z^2] - \underline{Z}^2$$

- Corollary: $E[Z^2] = E[\, (Z - \underline{Z})^2 \,] + \underline{Z}^2$

Bias-Variance-Noise Decomposition

$$E[(h(x^*) - y^*)^2] = E[h(x^*)^2 - 2\, h(x^*)\, y^* + y^{*2}]$$

$$= E[h(x^*)^2] - 2\, E[h(x^*)]\, E[y^*] + E[y^{*2}]$$

$$= E[(h(x^*) - \underline{h(x^*)})^2] + \underline{h(x^*)}^2 \quad \textit{(lemma)}$$

$$- 2\, \underline{h(x^*)}\, f(x^*)$$

$$+ E[(y^* - f(x^*))^2] + f(x^*)^2 \quad \textit{(lemma)}$$

$$= E[(h(x^*) - \underline{h(x^*)})^2] + \quad \textit{[variance]}$$

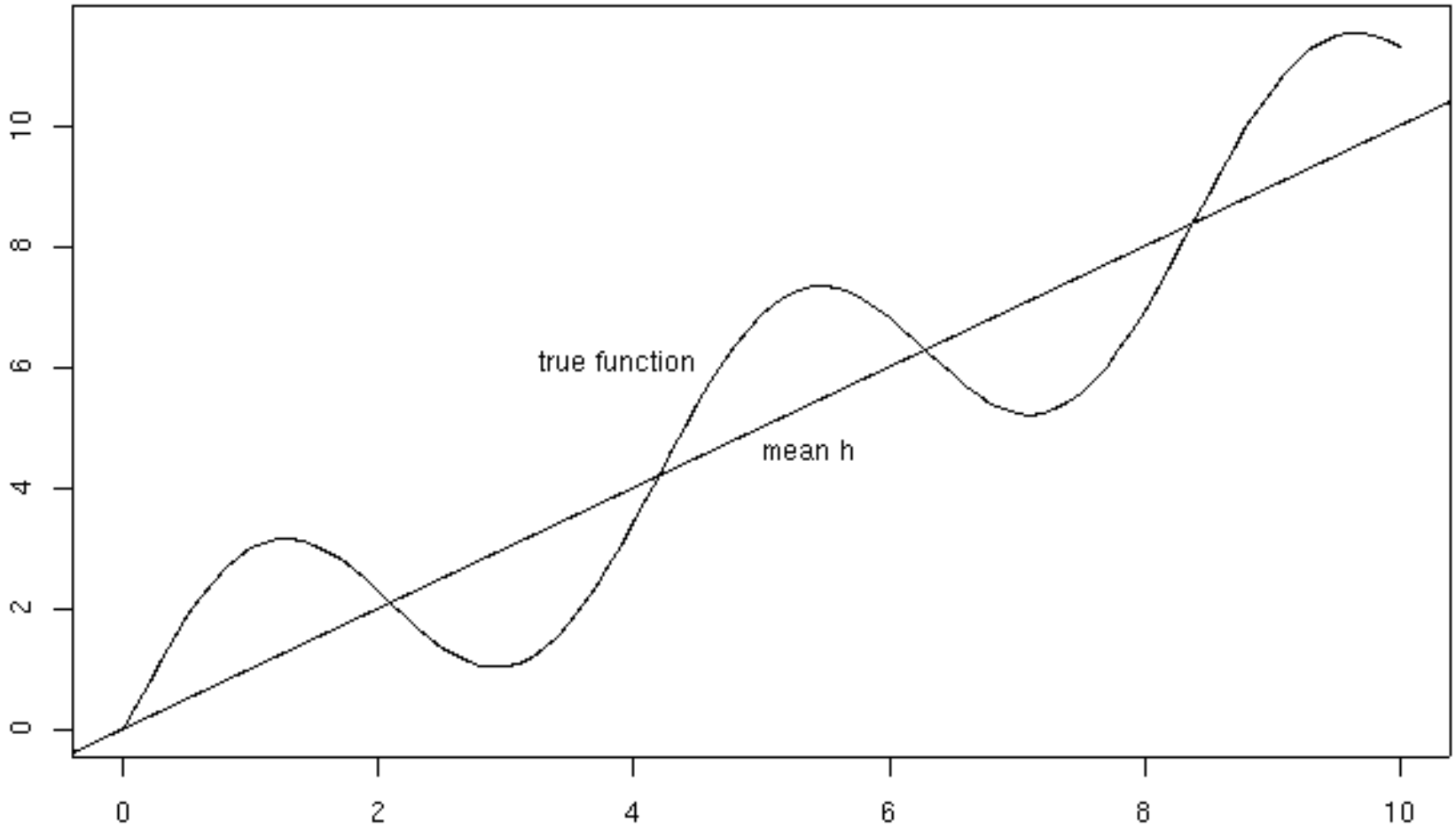
$$(\underline{h(x^*)} - f(x^*))^2 + \quad \textit{[bias}^2\textit{]}$$

$$E[(y^* - f(x^*))^2] \quad \textit{[noise]}$$

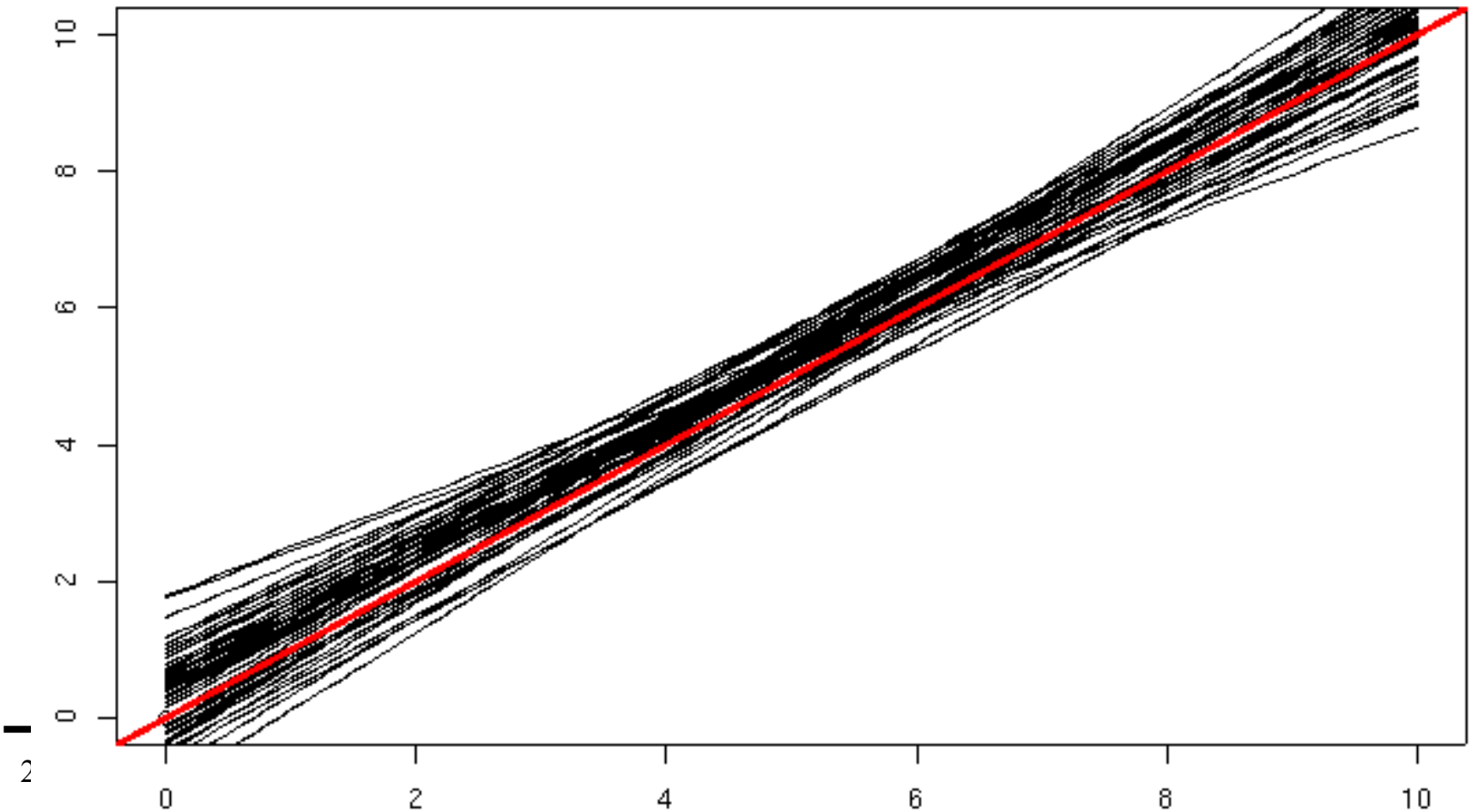
Bias, Variance, and Noise

- Variance: $E[\,(h(x^*) - \underline{h(x^*)})^2\,]$
  - Describes how much $h(x^*)$ varies from one training set S to another
- Bias: $[\underline{h(x^*)} - f(x^*)]$
  - Describes the <u>average</u> error of $h(x^*)$
- Noise: $E[\,(y^* - f(x^*))^2\,] = E[e^2] = s^2$
  - Describes how much y^* varies from $f(x^*)$

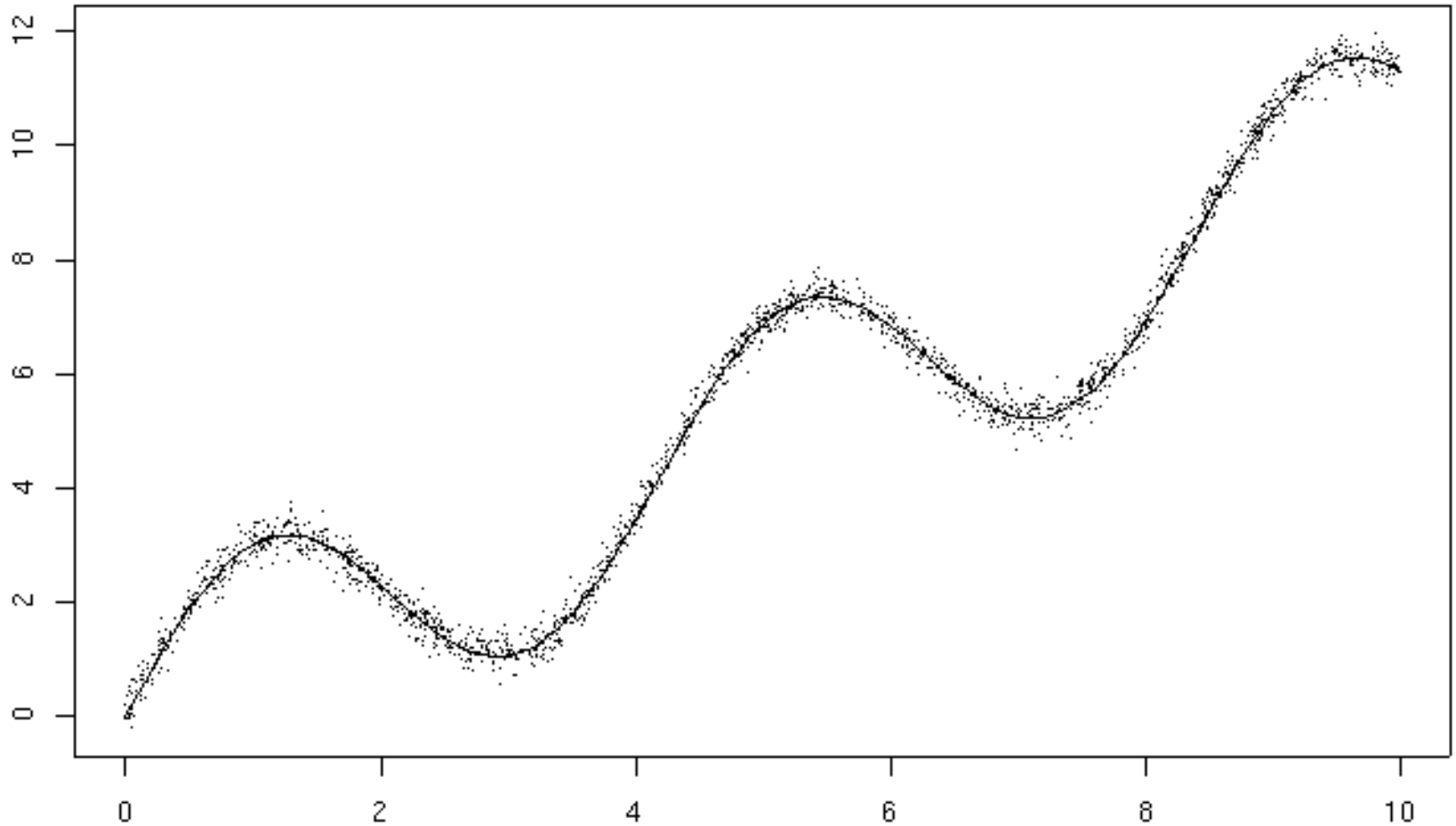
Bias



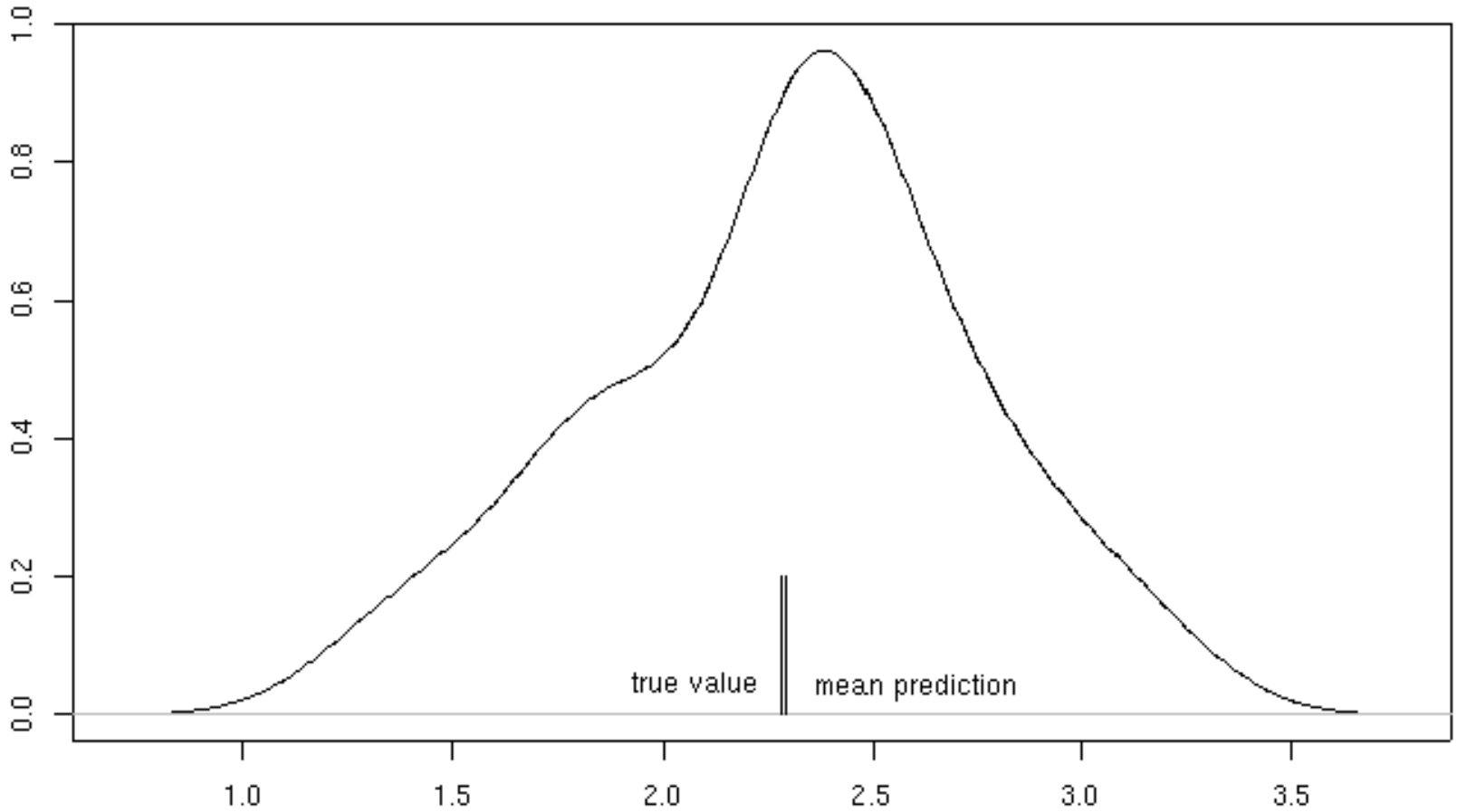
Variance



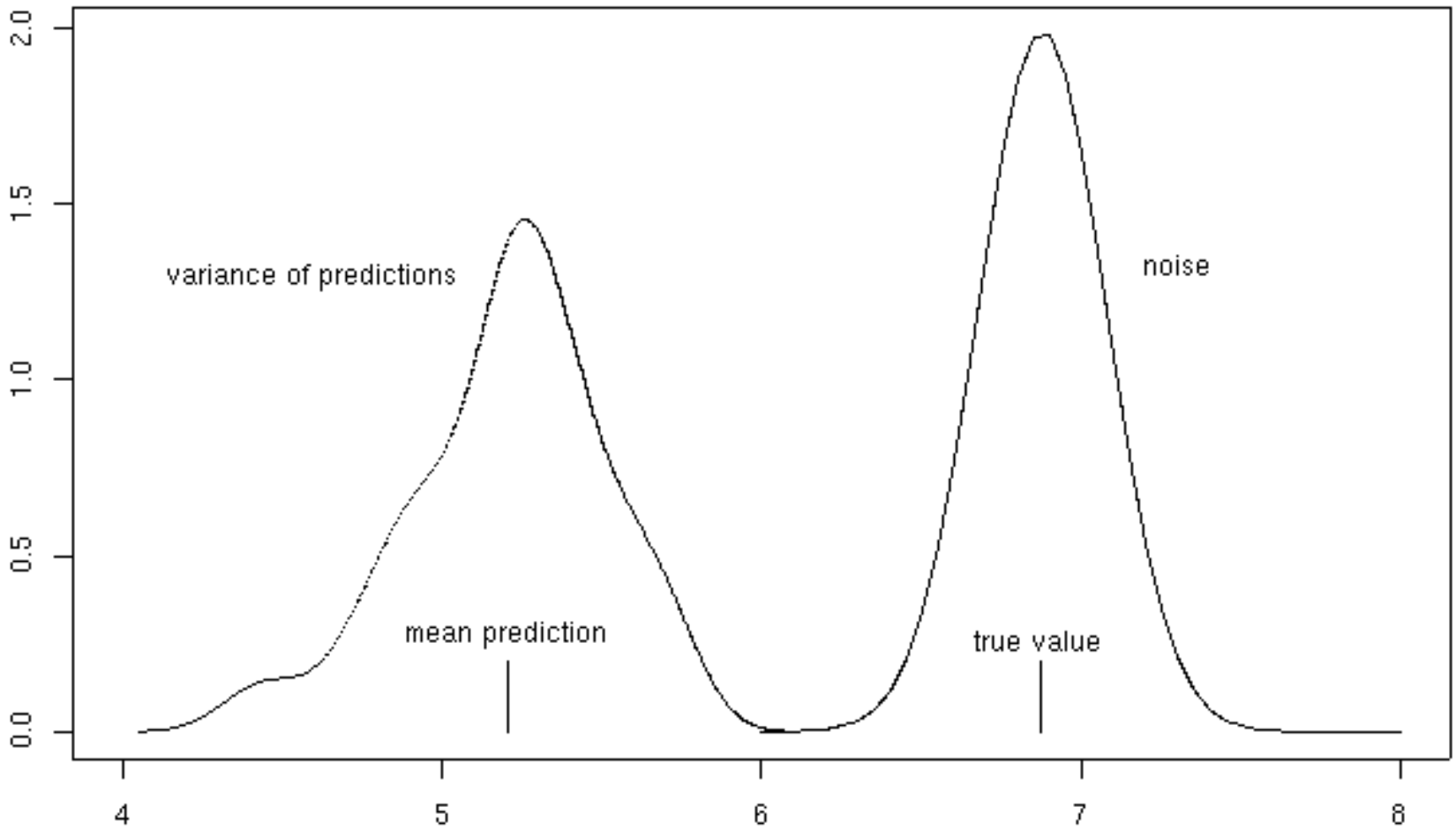
Noise



Distribution of Predictions at x=2.0



Distribution of Predictions at x=5.0



Measuring Bias and Variance

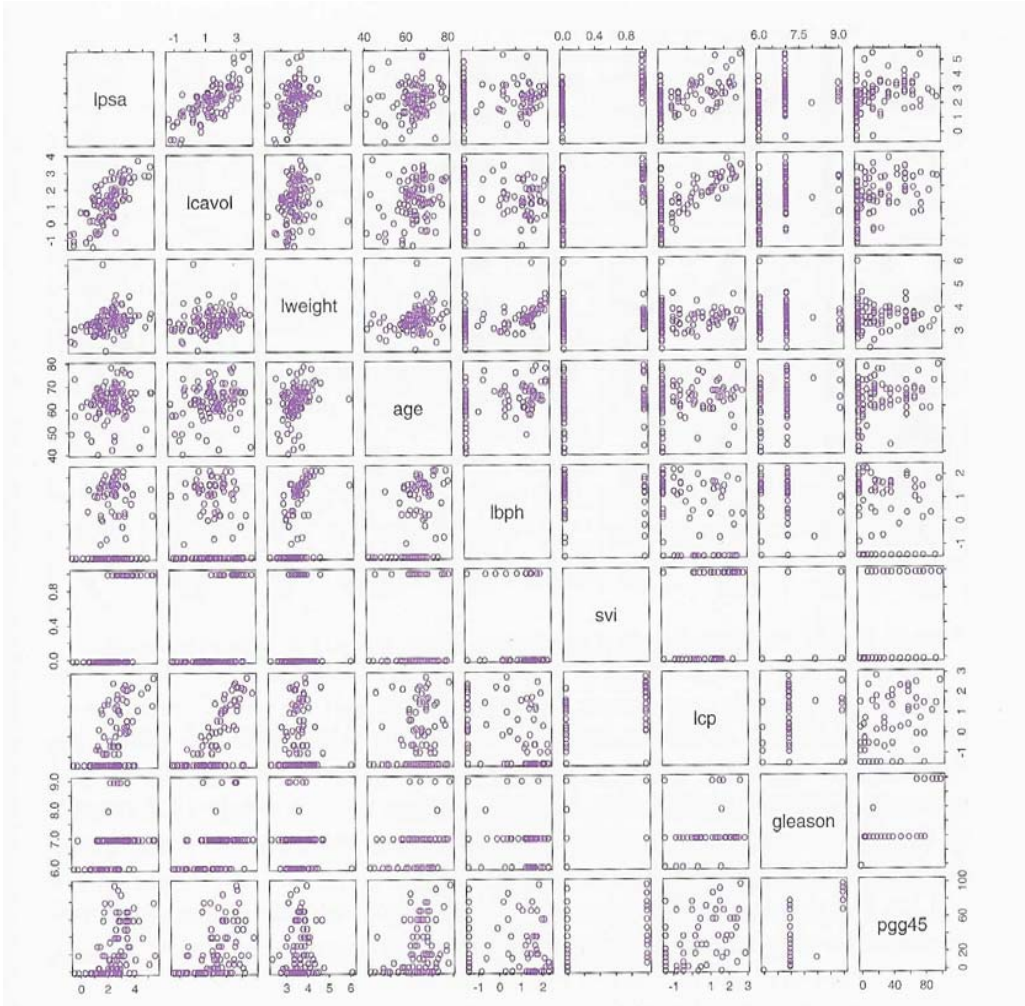
- In practice (unlike in theory), we have only ONE training set *S*
- We can simulate multiple training sets by bootstrap replicates
  - *S'* = {*x* | *x* is drawn at random with replacement from *S*} and |*S'*| = |*S*|

Estimating Bias and Variance

- For each run, at each data point x, we will now have the observed corresponding value y and several predictions $y_1, \ldots, y_K$
- Compute the average prediction $\underline{h}$
- Estimate bias as $(\underline{h} - y)$
- Estimate variance as $S_k=(y_k - \underline{h})^2/(K - 1)$
- Assume noise is 0

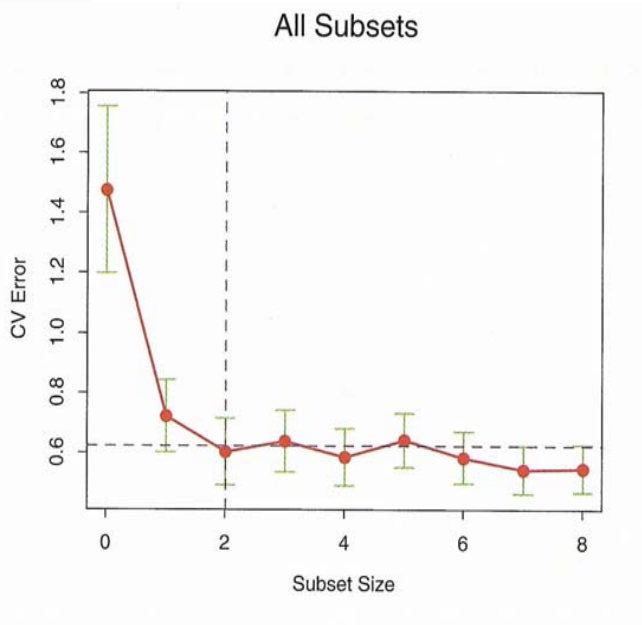
Prostate Cancer Data Example

- Variables :
  - Response: level of prostate-specific antigen
  - Regressors : 8 clinical measures useful for men receiving prostatectomy



Results

Term	LS	Best Subset	Ridge	Lasso	PCR	PLS
Intercept	2.480	2.495	2.467	2.477	2.513	2.452
lcavol	0.680	0.740	0.389	0.545	0.544	0.440
lweight	0.305	0.367	0.238	0.237	0.337	0.351
age	-0.141		-0.029		-0.152	-0.017
lbph	0.210		0.159	0.098	0.213	0.248
svi	0.305		0.217	0.165	0.315	0.252
lcp	-0.288		0.026		-0.053	0.078
gleason	-0.021		0.042		0.230	0.003
pgg45	0.267		0.123	0.059	-0.053	0.080
Test Error	0.586	0.574	0.540	0.491	0.527	0.636
Std. Error	0.184	0.156	0.168	0.152	0.122	0.172



Cross Validation

- For each method, the best compromise model is given
- Quality of the model is measured by tenfold cross-validation:

$$CV_k = \frac{1}{m}\sum_{i=1}^{m}(y_{ki} - \hat{f}^{-k}(x_i))^2 \quad k = 1...10$$

  - The data set is divided in ten parts, the model estimated by removing one by one each part and calculating the *cross-validation prediction error* on the m removed data.
- The *test error* is the mean of the CV_k's, the *std error* their standard deviation
- On the graph, the points are test errors and the intervals are at +/- 1s of the test error
- The best model is the less complex model with a mean CV within +/-1s of the best mean CV

Subset Selection Methods

- **Retain only the subset of variables giving the best fit**
- **Best subset regression :**
  - All subset of k=1,2,…p regressors are tested and for each k the best model retained. Work for p<30 to 40. The chosen model will be based on a criterion making a tradeoff of bias and variance
- **Forward selection :**
  - Start with the intercept and add at each step the predictors that most improves the fit. The improvement is often measured with p-value of an F-test
- **Backward selection**
  - Start with the full model and removes one by one the worst regressor. Stops when all regressors have a significant effect. Can only be used when p<N.
- **Stepwise selection**
  - Combines forward and backward to decide at each step which variable to remove and/or to add.

Shrinkage Methods: Ridge Regression

- Shrink coefficients by imposing a penalty to their size

$$\hat{\beta}^{ridge} = min_{\beta}\left\{\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}\beta_j^2\right\},\ \lambda > 0$$

This is equivalent to $\hat{\beta}^{ridge} = min_{\beta}\left\{\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2\right\}$ **subject to** $\sum_{j=1}^{p}\beta_j^2 \le s$

- Solutions not equivalent under scaling of inputs, standardize before solving. with centered $x_i - \overline{x}$
- Intercept with penalization and estimated by $\hat{\beta}_0 = \overline{y}$

$$RSS(\lambda) = (y - X\beta)^t(y - X\beta) + \lambda\beta\beta^t \qquad \hat{\beta}^{ridge} = (X^tX + \lambda I)^{-1}X^ty$$

If inputs are orthogonal $\hat{\beta}^{ridge} = \gamma\hat{\beta}^{LS},\ 0 \le \gamma \le 1$ Is a function of λ

Singular Value Decomposition (SVD)

$$X = UDV^T \qquad D \text{ diagonal} \qquad d_1 \geq d_2 \geq \ldots \geq d_p$$

Writing in terms of SVD

Coordinates of y respect to U basis

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

$$\hat{y} = X\beta^{ridge} = X(XX^T+\lambda I)^{-1}X^T y = UD(D+\lambda I)^{-1}DU^T y$$

$$= \sum_{j=1}^{p} \mathbf{u}_j \frac{d_j^2}{d_j^2+\lambda} \mathbf{u}_j^T y$$

shrinked coordinates of y respect to U basis

Greater amount of shrinkage to basis vector having smaller d_j^2

Complexity parameter of the model:

$$df(\lambda) = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2+\lambda}$$

Effective degrees of freedom

Relationship with Principal Components

Writing in terms of SVD $X^T X = V D^2 V^T$ *Eigen decomposition*

Eigen vectors v_j **are the principal component directions of** X

Normalized principal component

$$z_j = X v_j = \mathbf{u}_j d_j \qquad Var(z_j) = Var(X v_j) = \frac{d_j^2}{N}$$

Principal components Normalized linear combinations of the columns of X

Small values d_j^2 correspond to directions of the column space of X having small variance

Ridge regression assumes that response will tend to vary the most in the directions of higher variance of the inputs and then it protects against high variability of *Y* in directions of small variability of *X*

Lasso Regression

$$\hat{\beta}^{Lasso} = min_{\beta}\left\{\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2\right\} \quad \textbf{subject to} \quad \sum_{j=1}^{p}|\beta_j| \leq s$$

Solutions: Non-linear with respect to y_i

$$s = \frac{t}{\sum_{j=1}^{p}|\hat{\beta}_j|} \quad \text{where} \quad \hat{\beta}_j = \hat{\beta}_j^{LS}$$

- If *t* is small enough some coefficients are 0, so it acts like a subset selection method
- If $t \geq \sum_1^p |\hat{\beta}_j|$ then $\hat{\beta}_j^{Lasso} = \hat{\beta}_j^{LS}$

***t* is chosen to minimize de estimation of the expected prediction error**

Principal Component Regression (PCR)

Use of linear combinations of PCs: $z_m = Xv_m,\ j = 1, \ldots, M,\ M \leq p$

Since the z_m *are orthogonal*:

$$y_i^{PCR} = \hat{y} + \sum_{m=1}^{M} \hat{\theta}_m z_m \qquad \beta^{PCR}(M) = \sum_{m=1}^{M} \hat{\theta}_m v_m \qquad \hat{\theta} = \frac{<z_m, y>}{<z_m, z_m>}$$

- If $M = p,\ \hat{\beta}^{PCR} = \hat{\beta}^{LS}$ since $Z = UD$ span the space column of *X*
- If $M \leq p$ ⇨ reduced regression.

- Ridge and PC regression operate via the principal components *X*
- Ridge regression shrinks the regression coefficients of the principal components depending on the size of the corresponding eigenvalue.
- PC regression discard the *(p-M)* smallest eigentvalues

Partial Least Squares (PLS)

- Use a set of linear combinations of the inputs also taking into account y

1. Standardize x_j , set $\hat{y}^{(0)} = \bar{y}$ $x_j^{(0)} = x_j$ $j = 1, \ldots, p$
2. For $m = 1, \ldots, p$

$$\mathbf{z}_m = \sum_{j=1}^{p} \hat{\phi}_{mj} \mathbf{x}_j^{(m-1)} \quad \hat{\phi}_{mj} = < \mathbf{x}_j^{(m-1)}, y > \qquad \hat{\theta} = \frac{<z_m, y>}{<z_m, z_m>}$$

$$\hat{\mathbf{y}}^{(m)} = \hat{\mathbf{y}}^{(m-1)} + \hat{\theta}_m z_m \quad \text{output}$$

- Orthogonalize $\mathbf{x}_j^{(m-1)}$ with respect to $\mathbf{z}_m$

$$\mathbf{x}_j^{(m)} = \mathbf{x}_j^{(m-1)} - [\frac{<\mathbf{z}_m, \mathbf{x}_j^{(m-1)}>}{<\mathbf{z}_m, \mathbf{z}_m>}]\mathbf{z}_m$$

- The coefficients on the original $\mathbf{x}_j$: $\hat{\beta}_j^{PLS}(m) = \sum_{l=1}^{m} \hat{\phi}_{lj} \hat{\theta}_l$

Comparing PCR with PLS

- If M<P then reduced regression, M=p ⇒ least squares
- Non linear function of y
- Look for directions with high variance **and** high correlation with y

Covariance matrix of X

PCR: $max\ Var(X\alpha), \left\{||\alpha|| = 1, v_l^t \mathbf{S}\alpha = 0, l = 1, \ldots, m-1\right\}$

$z_m = X\alpha$

non correlated with previous linear combinations

PLS: $max\ Corr^2(y, X\alpha)Var(X\alpha), \left\{||\alpha|| = 1, v_l^t \mathbf{S}\alpha = 0, l = 1, \ldots, m-1\right\}$

Dominates, then behaves similar to Ridge and PCR

If X is orthogonal then $\hat{\beta}_j^{PLS}(m) = \hat{\beta}_j^{LS}(m),\ m = 1$

Multiple Outcome: Correlation Outputs

- Selection and shrinkage methods: univariate individually or simultaneously to all ouputs
  - Ridge: apply ridge regression to each column of *Y*
- Canonical correlation analysis (CCA similar to PCA)
  - Sequence of uncorrelated combinations on *X* , $\mathbf{X}v_m$, and *Y*, $\mathbf{Y}u_m$ such that $Corr^2(\mathbf{Y}u_m, \mathbf{X}v_m)$ are maximized
  - Leading canonical responses are those linear combinations best predicted by the x's
  - CCA solution is computed via SVD of the cross-covariance $\mathbf{Y}^T\mathbf{X}/N$
    - Reduced-rank regression: in terms of $Cov(\epsilon) = \Sigma$ estimated by $\mathbf{Y}^T\mathbf{X}/N$

$$\hat{\mathbf{B}}(m) = argmin_{\{rank(B)=m\}} \sum_{i}^{N} (y_i - \mathbf{B}^T x_i)^T \Sigma^{-1} (y_i - \mathbf{B}^T x_i)$$

$$\hat{\mathbf{B}}^{rr}(m) = \hat{\mathbf{B}} U_m U_m^-$$

$$\hat{\mathbf{Y}}^{rr}(m) = \mathbf{X}(\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{YUU}^- = \mathbf{HYP}_m$$

Comparison

- PLS, PCR and Ridge regression behave similarly.
- Ridge shrinks all but shrinks more low variance directions
- PCR keeps higher variance ones and discard the rest.
- PLS shrinks low-variance directions but tends to inflate some of the higher variance ones, causing slightly larger prediction error compared to ridge
- Ridge is preferred for minimizing the prediction error and it shrinks smoothly
- If X is orthogonal, these apply simple transformations to the least square estimates. If X is not orthogonal, they converge to the least squares

Finding Residuals & Estimating Variance

- Residuals = differences between Y and the regression line (the fitted line)
- An unbiased estimate of s^2 is: [sum of squared residuals]/ (n-2)
- Degree of freedom is n-2 because two parameters were estimated
- [sum of squared residuals]/s^2 follows a chi-square

Hypothesis Testing for Slope

- Slope estimate β_1 is random
- It follows a normal distribution with mean equal to the true β_1 and the variance equal to $\sigma^2 / [n \, \mathrm{var}(X)]$
- Because σ^2 is unknown, we have to estimate from the data; the SE (standard error) of the slope estimate is equal to the squared root of the above $\hat{s}_e^2 = \mathrm{MSE}$

Test H_0: $\beta_1=0$: $t_{b_1} = \dfrac{\hat{b}_1 - 0}{\dfrac{\hat{s}_e}{\sqrt{S_{xx}}}}$ Reject H_0 if: $t_{b_1} > t_{n-2,1-a}$

(1-α)100% CI

$$\hat{b}_1 - t_{n-1,(1-\frac{a}{2})} \frac{\sqrt{\mathrm{MSE}}}{\sqrt{Sxx}} < b_1 < \hat{b}_1 + t_{n-1,(1-\frac{a}{2})} \frac{\sqrt{\mathrm{MSE}}}{\sqrt{Sxx}}$$

t-Distribution

- Suppose an estimate $\hat{\theta}$ is normal with variance σ^2
- Suppose σ^2 is estimated by s^2 which is related to a chi-squared distribution
- Then $(\hat{\theta} - \theta)/ (s^2)$ follows a *t*-distribution with the degrees of freedom equal to the chi-square ~~degree freedom~~

Model Selection

- Cross-validation: Measure generalization accuracy by testing on data unused during training
- Regularization: Penalize complex models

  E'=error on data + λ model complexity

- Akaike's information criterion (AIC), Bayesian information criterion (BIC)
- Minimum description length (MDL): Kolmogorov complexity, shortest description of data
- Structural risk minimization (SRM): in SVM

Summary

- Today's Class
  - Linear method for regression (Chapter 3)
- Next Class
  - Autoregression and linear predictive speech analysis
  - Linear methods for classification
- Exercises: make sure you know the topics discussed and how to do all the exercises suggested in Lecture 2
- Reading Assignments
  - HTF, Chapter 4