Day 4, 9/18/18 **Weighted regression**

Each observation is characterized by different variability

Example: each obs. is an average of different number of points

Recall: for iid Y_i, $Var\left(\frac{1}{N}\sum_{i=1}^{N} Y_i\right) = \frac{1}{n} Var(Y_i)$

In linear regression: $Y = X\beta + \varepsilon$, $Var(\varepsilon) = \sigma^2 \Sigma$, $\Sigma = \begin{pmatrix} \frac{1}{n_1} & & \\ & \frac{1}{n_2} & \\ & & \ddots \frac{1}{n_N} \end{pmatrix}$

Change of notation: $Y = X\beta + \varepsilon$, $Var(\varepsilon) = \sigma^2 W^{-1}$, $W = \begin{pmatrix} n_1 & & \\ & n_2 & \\ & & \ddots n_N \end{pmatrix}$

Linear transformation: $\underbrace{W^{1/2}Y}_{Y_W} = \underbrace{W^{1/2}X}_{X_W}\beta + \underbrace{W^{1/2}\varepsilon}_{\varepsilon_W}$

$Var(\varepsilon_W) = Var(W^{1/2}\varepsilon) = \sigma^2 W^{1/2}\Sigma W^{1/2} = \sigma^2 \Sigma W^{1/2}W^{1/2} = \sigma^2 I$

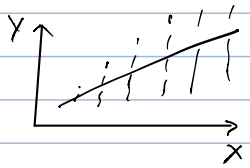
Change in loss function

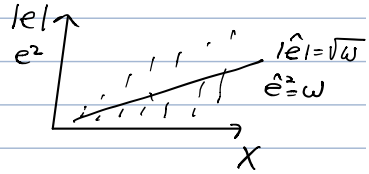
Minimizing squared error on the transformed scale:

$$\min_{\beta} (Y_W - X_W\beta)^T (Y_W - X_W\beta) = (W^{1/2}Y - W^{1/2}X\beta)^T (W^{1/2}Y - W^{1/2}X\beta) = (Y - X\beta)^T W (Y - X\beta) = \sum_{i=1}^{N} w_i (Y_i - X_i^T\beta)^2$$

(X_i: $1 \times p$)

$\Rightarrow \hat{\beta} = (X_W^T X_W)^{-1} X_W^T Y_W = (XWX)^{-1} X'WY$ — averages with larger n_i have more weight

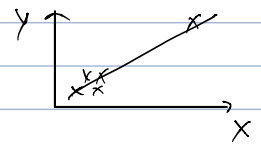




Example $Var(Y_i | X_i) = f(X_i)$

Approach:
(1) Fit OLS regression
(2) Find a relationship between $|e_i|$ or e_i^2 and X_i
(3) Use the relationship to quantify weights
(4) Refit regression with the weights
(5) Optional: iterate (1)-(4)

Observations with high leverage



Also outliers — see hw 1

Recall $\hat{Y} = \underbrace{X(X'X)^{-1}X^T}_{H:\ \text{hat matrix}} Y$ ($\hat{\beta}$)

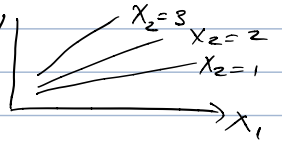
Leverage statistic for obs. i: ith diagonal element of H.

2-variable regression: $h_i = \frac{1}{N} + \frac{(X_i - \bar{X})^2}{\sum_{j=1}^{N}(X_j - \bar{X})^2}$

Expressing non-linearity

Polynomial regression: $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots \beta_{p-1} X^{p-1} + \varepsilon$

Regression with higher order terms: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$



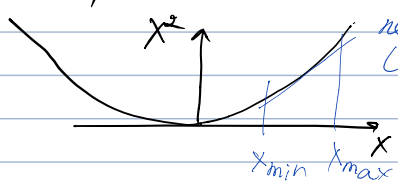
- surface with curvature

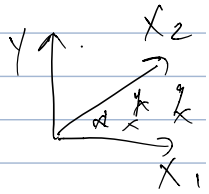
Price for more flexible models: more parameters

Multicollinearity

$\hat{\delta} = (X^TX)^{-1}X^TY$ cannot invert X^TX when $\text{rank}(X) < p$
- X has linearly dependent columns
- $N < p$

Example: higher order terms
Solution: center and scale





Bias - variance tradeoff

Until now assumed that the linear model is true. How to select the right model? what is the price for the wrong model?

Consider loss functions again:

$J(\delta) = \sum_{i=1}^{N} (y_i - \hat{h}_\delta(x_i))^2$ - squared loss/squared error : evaluate $h_\delta(x)$ on current data

Generalization error: performance on future, independent test data
= prediction error

$\text{Error}_{\{x_i,y_i\}} = E[J(\delta)|\{x_i,y_i\}]$ - expected loss, ave. over all future data, given training set $\{x_i,y_i\}$ - more meaningful

$\text{Error} = E[J(\hat{\delta})]$ - expected loss, ave. over all future data and all training sets
- easier mathematically

Prediction for $x = x_0$ squared loss

$$\text{Error}(x_0) = E[Y - \hat{h}_\delta(x_0)]^2 = E[Y - h(x_0) + h(x_0) - E[\hat{h}_\delta(x_0)] + E[\hat{h}_\delta(x_0)] - \hat{h}_\delta(x_0)]^2$$

$$= \underbrace{E[Y - h(x_0)]^2}_{\text{irreducible error}} + \underbrace{E[h(x_0) - E\{\hat{h}_\delta(x_0)\}]^2}_{\text{bias}^2} + \underbrace{E[E\{\hat{h}_\delta(x_0)\} - \hat{h}_\delta(x_0)]^2}_{\text{variance}^2}$$

- cross-products cancel out

bias^2: model mis-specification

variance^2: variance of the sampling distribution of parameter estimates

[Ave model bias]^2 [Ave estimation bias]^2

Specific to linear regression:

$$\text{Error}(x_0) = \sigma_\varepsilon^2 + E[h(x_0) - E\{\hat{h}_\delta(x_0)\}]^2 + \underbrace{\| X(X^TX)^{-1}x_0 \|_2^2 \sigma_\varepsilon^2}$$

$$\to E[E[x_0^T(X^TX)^{-1}X^TY] - x_0^T(X^TX)^{-1}X^TY]^2$$

$$= E\{[x_0^T(X^TX)^{-1}X^T(E\{y\} - y)]^2\}$$

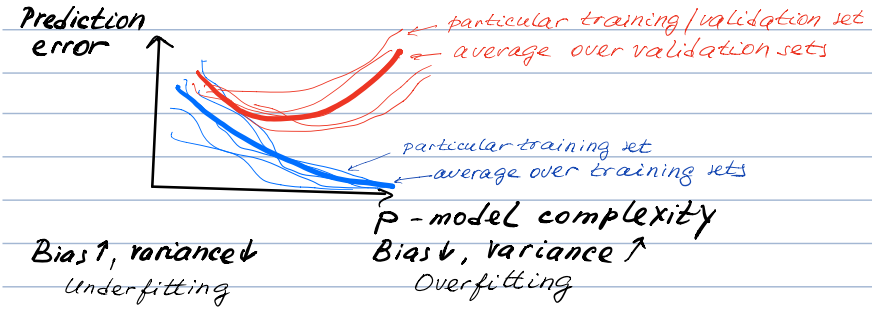
$$= [x_0^T(X'X)^{-1}X^T]\,\text{Var}(y)\,[x_0^T(X'X)^{-1}X^T]'$$

$$= \| X(X'X)^{-1}x_0 \|_2 \sigma_\varepsilon^2$$

How to estimate the error in practice? Average over future data

$$\frac{1}{N}\sum_{i=1}^{N} \text{Error}(x_0) = \sigma_\varepsilon^2 + \frac{1}{N}\sum_{i=1}^{N}\left[h(x_0) - E\{\hat{h}_0(x_0)\}\right]^2 + \frac{p}{N}\sigma_\varepsilon^2$$

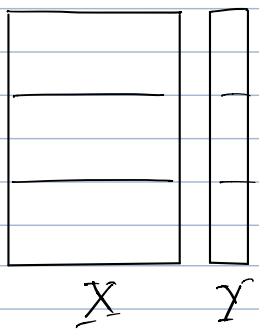
Training data	$p \uparrow \Rightarrow$	$\downarrow$	$\downarrow$	$\downarrow$	- apparent (in-sample) estimates
Future data	$p \uparrow \Rightarrow$	—	$\downarrow$	$\uparrow$	- true (independent) estimates



Goal: minimize prediction error on the validation set

In linear regression, model selection / model building = variable selection

Steps of variable selection $N \gg p$



- model fitting / training
- model development
- model evaluation

(1) Explore the space of candidate models
(2) Evaluate the candidate models on the dev. set, keep the best performing
(3) Evaluate the selected model on the evaluation set

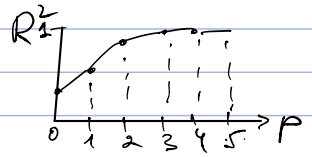
Exploring the model space in (1):
- Exhaustive search of $\binom{K}{p}$ predictors
- Forward stepwise selection
- Backward stepwise selection

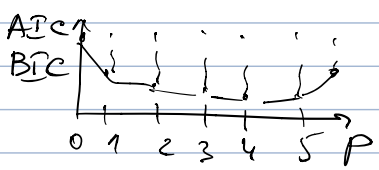
] many heuristics exist

Example: forward stepwise selection
- Denote M_0 the null model with no predictors
- For $k = 0, \ldots, p-1$:
  - consider all $p-k$ models that augment M_k with one predictor
  - $M_k \leftarrow$ the best among $p-k$ models

] R^2, AIC, BIC

- Select the model among $M_0, \ldots, M_p$ that performs best on dev set





$$AIC = \frac{SSE}{MSE_p} + 2p$$

$$BIC = \frac{SSE}{MSE_p} + \log N\, p$$

For Normal distribution $\Rightarrow$ AIC and BIC

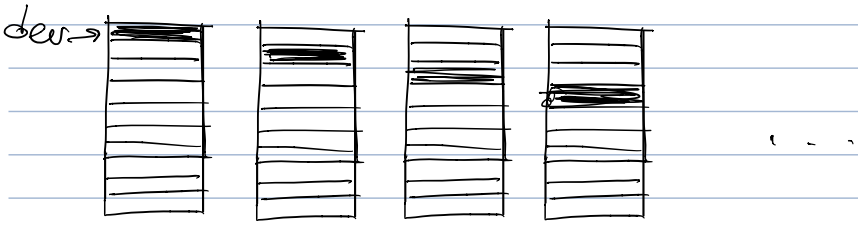
$$-2 \log L = \frac{\sum_{i=1}^{N} (x_i - \hat{x})^2}{\hat{\sigma}^2}$$

$\hat{\sigma}^2$ estimates by variance from the saturated model

are penalized log-L criteria

Data-poor situation: cross-validation $N \not\gg p$

Iterate partitioning into train and dev sets



Problem: each iteration does not always have the same best model.

"Consensus" model may not perform as well

Frequently used as an intermediate parameter tunning step