Seoul National University
Dept. of Civil and Environmental Engineering

Instructor: Junho Song
junhosong@snu.ac.kr

M1586.002500 Information Engineering for Civil & Environmental Engineers
In-Class Material: Class 07
Linear Regression (ISL Chapter 3)

1. Extensions of Linear Regression Model

Two important restrictive assumptions of the linear regression:

- '**Additive**' assumption: the effect of changes in X_j on the response Y is i________ of the values of the other predictors
- '**Linear**' assumption: the change in the response Y due to a one-unit change in X_j is c________, regardless of the value of X_j

(a) Can we remove the additive assumption by modifying the linear regression model?

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

One way of extending this model to allow for i__________ effects is to include a third predictor, called an **interaction term**, i.e.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

Rewritten as

$$\mathrm{Y} = \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \epsilon = \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 + \epsilon$$

where, $\tilde{\beta}_1 = \beta_1 + \beta_3 X_2$

$\tilde{\beta}_1$ changes with X_2, $\rightarrow$ the effect of X_1 on Y is no longer c________

Note: The hierarchical principle states that if we include an interaction term ($X_1 X_2$) in a model, we should also include the main effects (X_1 or X_2), even if the p-values associated with their coefficients are not significant. It does not make sense to talk about interaction effect while ignoring that of the predictors.

```
library(MASS) # Boston data in MASS
lm.fit1 = lm(medv~lstat+age, data=Boston)
lm.fit2 = lm(medv~lstat*age, data=Boston)
summary(lm.fit1)
summary(lm.fit2) #compare r.squared and RSE to see interaction effects
```

```
# install.packages("ISLR")
library(ISLR) # Carseats data in ISLR
summary(Carseats) # Car seats sales data at 400 stores (see ShelveLoc)
attach(Carseats)
contrasts(ShelveLoc) # dummy variables introduced for ShelveLoc

lm.fit3 = lm(Sales ~ . + Income:Advertising + Price:Age, data=Carseats)
summary(lm.fit3)
```

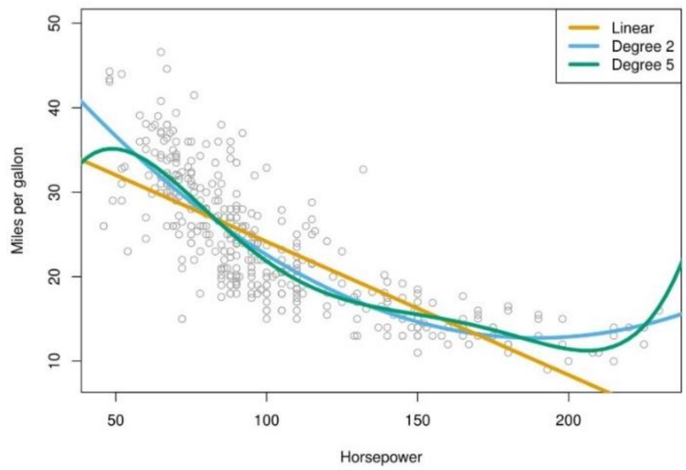
Instructor: Junho Song
junhosong@snu.ac.kr

(b) What about linear assumption? Can we remove it by modifying linear regression too?

→ **The p_________ regression** is a simple extension of the linear relationship between response and predictor to nonlinear one

Comparison between 1st, 2nd and 5-th order regression models of mpg with respect to horsepower in **Auto** data set, what can be inferred from this?

$$\begin{cases} \text{mpg} = \beta_0 + \beta_1 \text{horsepower} + \epsilon \\ \text{mpg} = \beta_0 + \beta_1 \text{horsepower} + \beta_2 \text{horsepower}^2 + \epsilon \\ \text{mpg} = \beta_0 + \beta_1 \text{horsepower} + \cdots + \beta_5 \text{horsepower}^5 + \epsilon \end{cases}$$



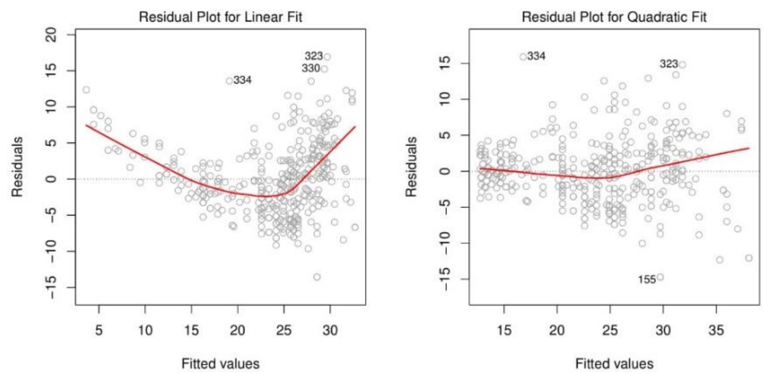

```
attach(Boston)
lm.fit_p1 = lm(medv ~ lstat, data=Boston)
lm.fit_p2 = lm(medv ~ lstat + I(lstat^2), data=Boston)
summary(lm.fit_p1)
summary(lm.fit_p2)

anova(lm.fit_p1, lm.fit_p2)
# ANOVA: analysis of variance
# if p value is small, the second model is significantly better
# https://bookdown.org/ndphillips/YaRrr/comparing-regression-models-with-
   anova.html

lm.fit_p5 = lm(medv ~ poly(lstat,5))
# 5-degree polynomial regression model
anova(lm.fit_p2,lm.fit_p5)
summary(lm.fit_p1)$r.squared
summary(lm.fit_p2)$r.squared
summary(lm.fit_p5)$r.squared
```

2. Potential Problems regarding Linear Regression

(a) **Non-linearity** of the response-predictor relationships

- [Problem] The basic assumption of the model is not satisfied
  → All of the conclusions drew from the fit are suspicious

- [Diagnosis] **Residual plots** (simple: $e_i = y_i - \hat{y}_i$ versus x_i, multiple: e_i versus $\hat{y}_i$)
  Ideally, the residual plot should show no discernible pattern

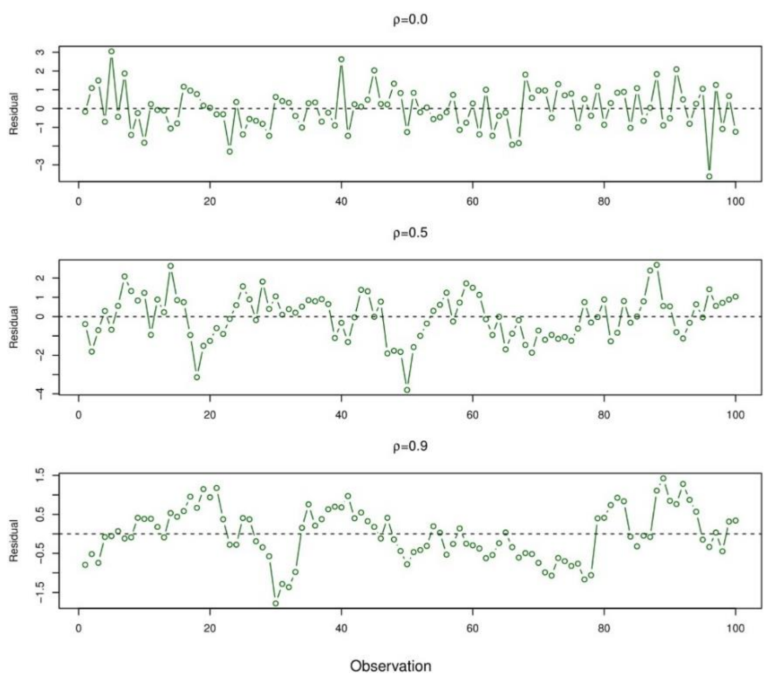


Residual plots versus fitted variables from **Auto** data set

- Left: a strong pattern in the residuals indicates non-linearity in the relationship
- [Solution] Using non-linear transformations of the predictors, e.g. $\log X, \sqrt{X}, X^2$
- Right: little pattern in the residuals → quadratic term improves the fit to the data

(b) **Correlation** of error terms

- [Problem] An assumption that error terms $\epsilon_1, \epsilon_2, \cdots, \epsilon_n$ are uncorrelated is not satisfied
  → estimated standard errors tend to underestimate the true standard errors
  → confidence and prediction intervals will be narrower than the actual ones
- [Diagnosis] Usually occurs in time series problem. In residuals plot as a function of time, if the error terms are positively correlated, then tracking is observed
- [Solution] Many methods exist that take account of correlations in the error terms



Residual plots with different levels of correlation ρ between error terms

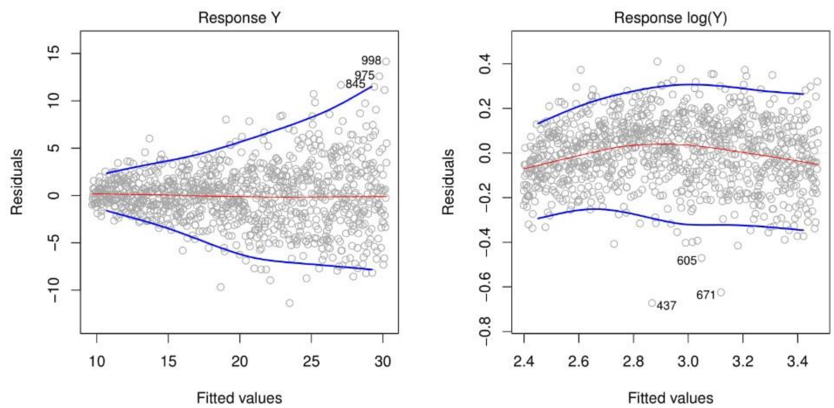
Instructor: Junho Song
junhosong@snu.ac.kr

- Top panel: no evidence of a time-related trend in the residuals
- Bottom panel: a clear pattern in the residuals; adjacent residuals tend to take on similar values.

(c) **Non-constant variance** of error terms

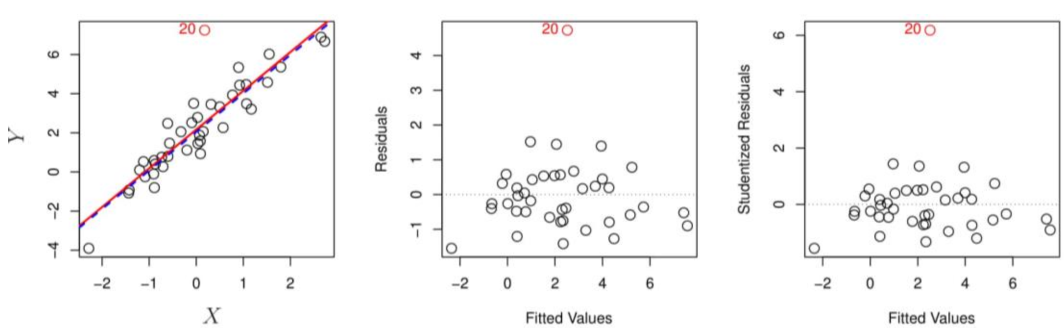
- [Problem] The assumption that error terms have constant variance, i.e. $\mathrm{Var}(\epsilon_i) = \sigma^2$, is not satisfied
- [Diagnosis] In residual plot, non-constant variances in the errors (heteroscedasticity), from the presence of a *funnel shape* in the residual plot.
- [Solution] Transform the response Y using a concave function such as $\log Y$ or $\sqrt{Y}$ → results in a greater amount of shrinkage of the larger responses

  What else? e.g. weighted least squares (larger weight on samples with smaller residuals)



(d) **Outliers**: points at which y_i is far from the value predicted by the model

- [Problem] Even if an outlier does not have much effect on the least squares fit, it can cause dramatic increase in RSE and decrease in R^2
- [Diagnosis] Residual plot can be one way to identify clear outliers
  In practice, a plot of "**studentized residuals**", i.e. ϵ_i/SE, is used
  → Observations whose studentized residuals are greater than 3 in absolute value: possible outliers



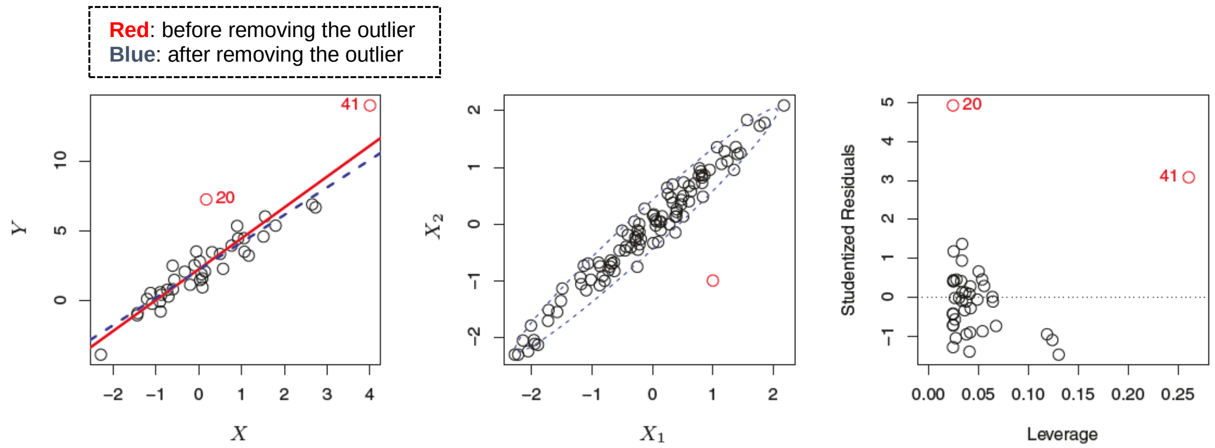
Red: before removing the outlier
Blue: after removing the outlier

Instructor: Junho Song
junhosong@snu.ac.kr

- [Solution] Remove outliers but with a caution (might indicate a deficiency with the model, e.g. a missing predictor)

(e) **High-leverage** points: unusual value for predictor x_i

- [Problem] It can cause a sizable impact on the estimated regression line. For this reason, it is important to identify high leverage observations



- [Diagnosis] Leverage statistic is used to quantify an observation's leverage value, for a simple linear regression

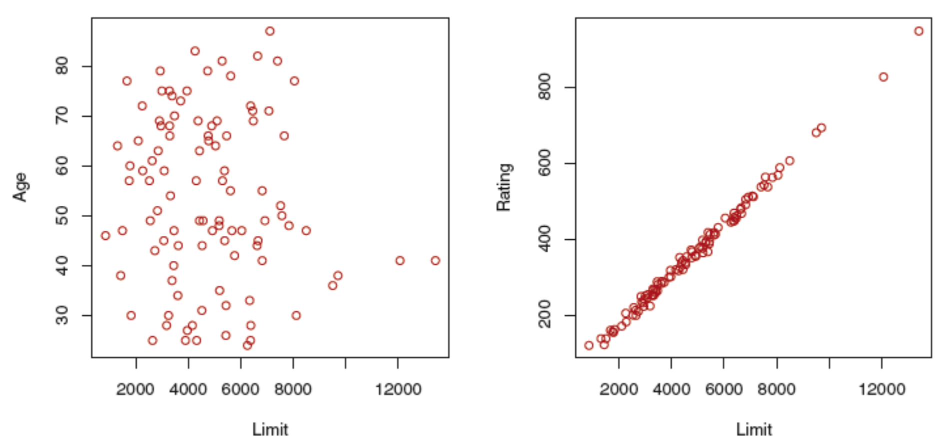
$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$$

Note: $\frac{1}{n} \leq h_i \leq 1,\ \ average\ leverage = \frac{p+1}{n}$

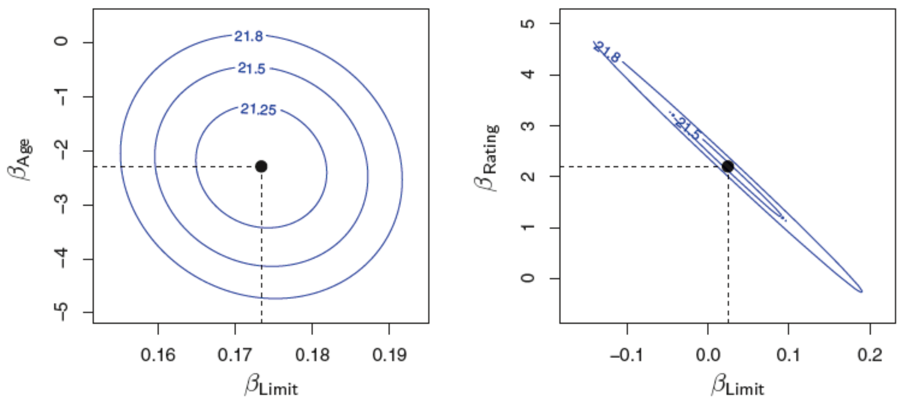
- [Solution] Remove predictor x_i that has significantly larger leverage than average

(f) **Collinearity**: Two or more predictor variables are closely related to one another

- [Problem] It can pose problems, since it is difficult to determine how each one is separately associated with the response



Scatter plots of predictors: not collinear (left), and highly collinear (right)



Contours of RSS: not collinear (left), and highly collinear (right)

If predictors are highly collinear, a small change in the data could cause the least squares estimates to move anywhere along narrow valley. This results in a great deal of uncertainty in the coefficient estimates.

Collinearity reduces the accuracy of the estimates of the regression coefficients, it causes the standard error for $\hat{\beta}_j$ to grow and the absolute value of the t-statistic to decrease.

		Coefficient	Std. error	t-statistic	p-value
Model 1	`Intercept`	−173.411	43.828	−3.957	< 0.0001
	`age`	−2.292	0.672	−3.407	0.0007
	`limit`	0.173	0.005	34.496	< 0.0001
Model 2	`Intercept`	−377.537	45.254	−8.343	< 0.0001
	`rating`	2.202	0.952	2.312	0.0213
	`limit`	0.025	0.064	0.384	0.7012

- [Diagnosis 1] Look at the correlation matrix (simple way)
  → Estimated standard errors tend to underestimate the true standard errors. This can be detected by inspection of the correlation matrix

Multicollinearity: It is possible for collinearity to exist between three or more variables even if no pair of variables has a particularly high correlation.

- [Diagnosis 2] Compute the variance inflation factor (VIF) to assess the multicollinearity. It is the variance of $\hat{\beta}_j$ when fitting the full model divided by the variance of $\hat{\beta}_j$ if fit on its own (smallest value 1, i.e. no collinearity):

$$\mathrm{VIF}(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}}$$

where $R^2_{X_j|X_{-j}}$ is the R^2 from a regression of X_j onto all of the other predictors

Instructor: Junho Song
junhosong@snu.ac.kr

$R^2_{X_j|X_{-j}} \approx 1$ → Collinearity is present → Large VIF

- [Solution] Drop one of the problematic variables (>5 or 10) from the regression or combine the collinear variables together into a single predictor

```
library(MASS) # Boston data in MASS
lm.fit = lm(medv~., data=Boston)
install.packages('car') # "Companion to Applied Regression" package
install.packages('cellranger') #should install 'cellranger' package before
   loading 'car' library
library(car)
vif(lm.fit) # compute variance inflation factor
```

3. K-Nearest Neighbors Regression (KNN Regression)

(a) Parametric VS Non-parametric method

Parametric method (e.g. Linear regression)

- Easy to fit, small number of coefficients, simple interpretation, need strong assumptions, if assumptions is far from truth, poor performance

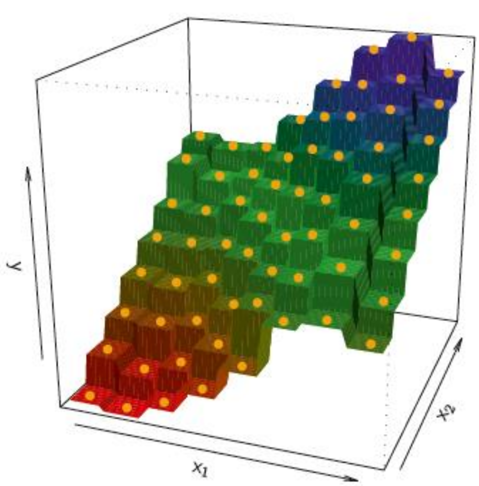
Non-parametric method (e.g. KNN regression)

- Do not need assumptions, more flexible, complex interpretation

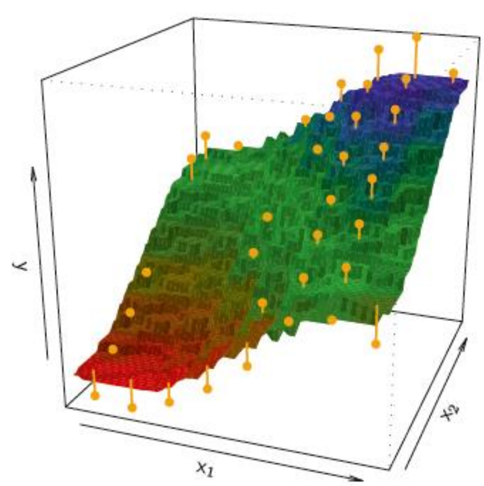
(b) Estimation of KNN regression

Given K and a prediction point x_0, identify the K training observations that are closest to x_0, represented by $\mathcal{N}_0$ and use the average of all the training responses in $\mathcal{N}_0$

$$\hat{f}(x_0) = \frac{1}{K} \sum_{x_i \in \mathcal{N}_0} y_i$$



K=1



K=9

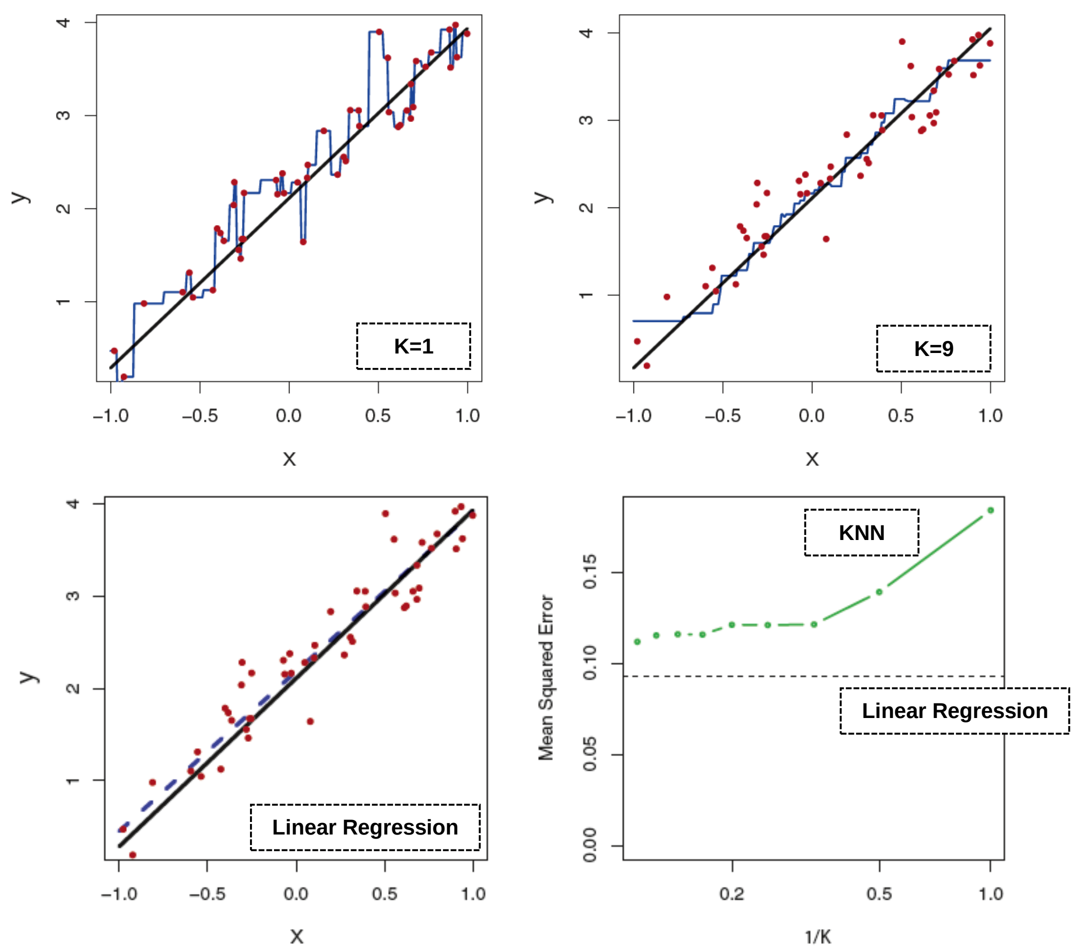
- Small K: Flexible fit   Bias $\downarrow$ Variance $\uparrow$
- Large K: Smooth fit   Bias $\uparrow$ Variance $\downarrow$

Optimal value for K depends on the **bias-variance tradeoff**. Recall (Class 03)

$$\mathrm{E}\left[\left(y-\hat{f}\right)^2\right] = \left[\mathrm{Bias}(\hat{f})\right]^2 + \mathrm{Var}(\hat{f}) + \mathrm{Var}(\epsilon)$$

(c) Comparison of linear regression with KNN regression

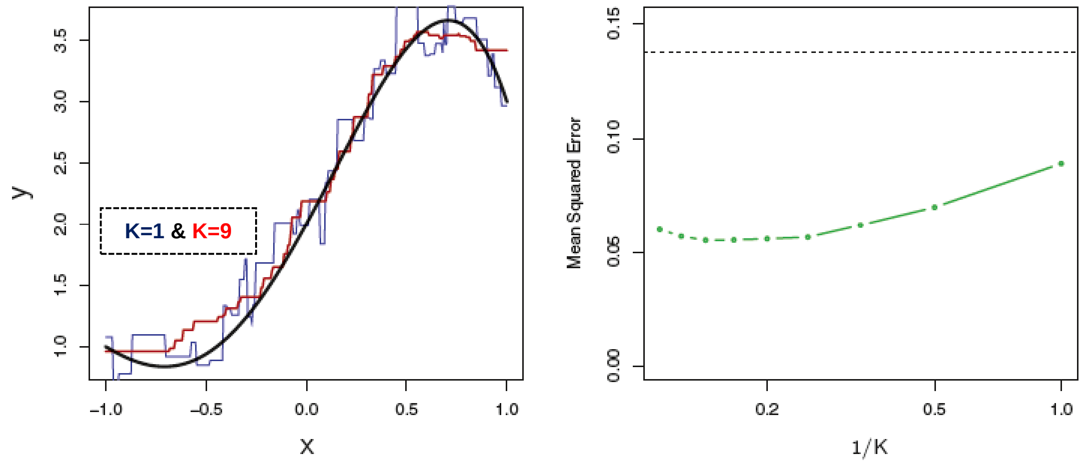
Case 1) True relationship between X and Y is linear



The parametric approach outperforms the nonparametric approach

Instructor: Junho Song
junhosong@snu.ac.kr

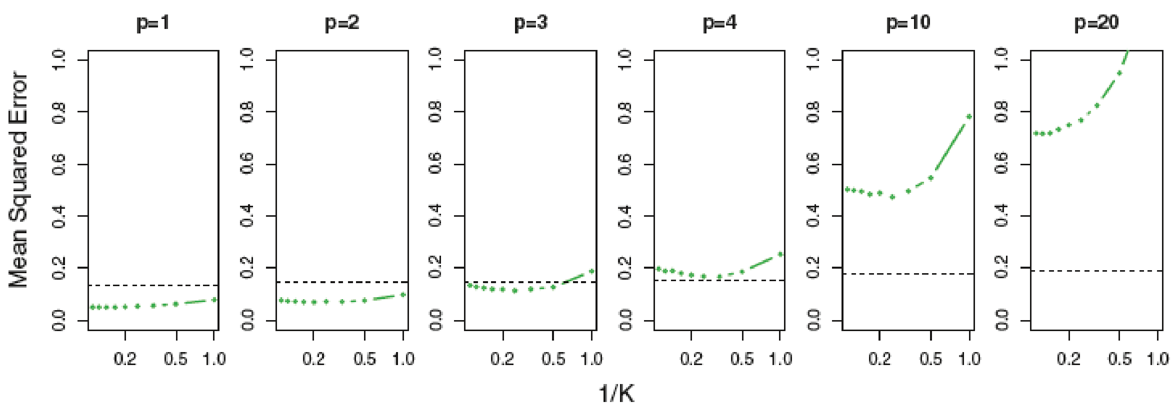
Case 2) True relationship between X and Y is strongly non-linear



The nonparametric approach outperforms the parametric approach

In general, a parametric model is superior to nonparametric model if the assumptions are valid.

(d) Curse of dimensionality (in KNN regression)



In high-dimensional problems, a given observation x_0 may have no nearby neighbors → "curse of dimensionality" (i.e. sparse coverage of the high-dimensional predictor space by data) → leading to a poor prediction of $f(x_0)$

Seoul National University
Dept. of Civil and Environmental Engineering

Instructor: Junho Song
junhosong@snu.ac.kr

M1586.002500 Information Engineering for Civil & Environmental Engineers
In-Class Material: Class 08
Linear Regression (ISL Chapter 3)

1. K-Nearest Neighbors Regression (KNN Regression)

(a) Parametric VS Non-parametric method

Parametric method (e.g. Linear regression)

- Easy to fit, small number of coefficients, simple interpretation, need strong assumptions, if assumptions is far from truth, poor performance

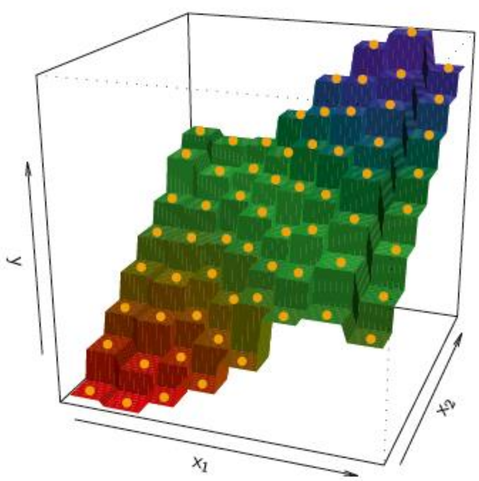
Non-parametric method (e.g. KNN regression)

- Do not need assumptions, more flexible, complex interpretation

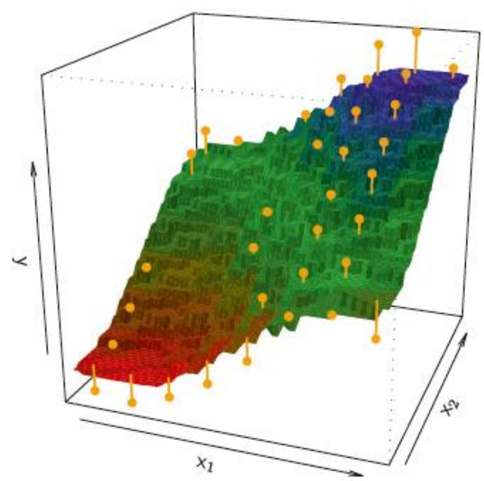
(b) Estimation of KNN regression

Given *K* and a prediction point x_0, identify the *K* training observations that are closest to x_0, represented by $\mathcal{N}_0$ and use the average of all the training responses in $\mathcal{N}_0$

$$\hat{f}(x_0) = \frac{1}{K}\sum_{x_i\in\mathcal{N}_0} y_i$$



K=1



K=9

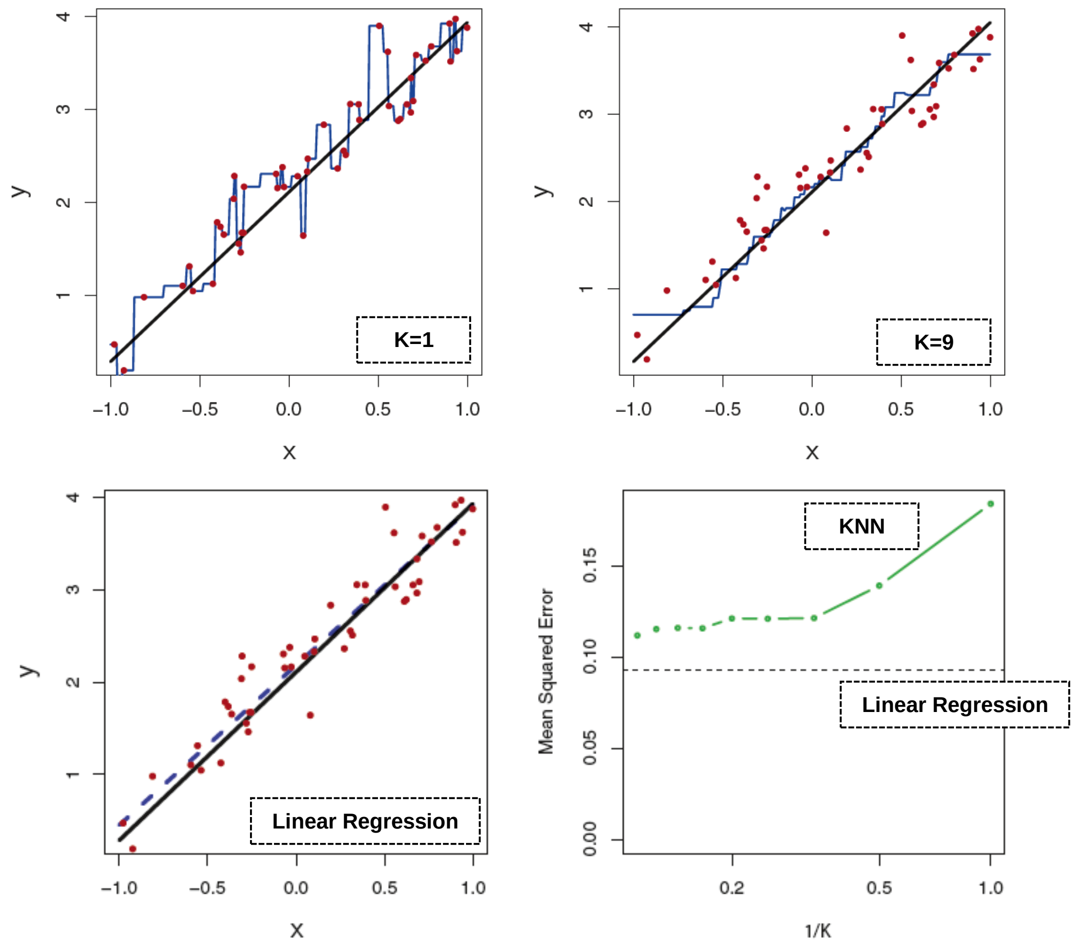
- Small K: Flexible fit Bias ↓ Variance ↑
- Large K: Smooth fit Bias ↑ Variance ↓

Optimal value for K depends on the **bias-variance tradeoff**. Recall (Class 03)

$$\mathrm{E}\left[\left(y-\hat{f}\right)^2\right] = \left[\mathrm{Bias}(\hat{f})\right]^2 + \mathrm{Var}(\hat{f}) + \mathrm{Var}(\epsilon)$$

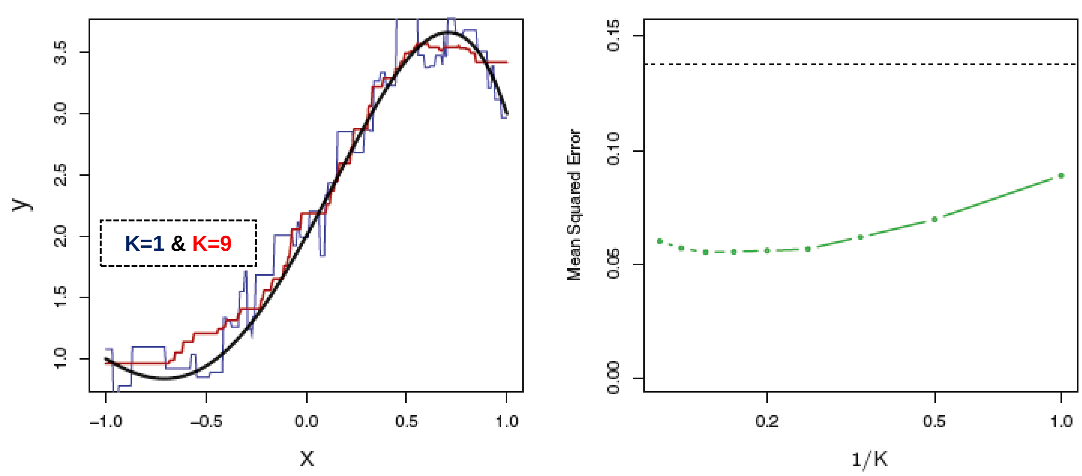
(c) Comparison of linear regression with KNN regression

Case 1) True relationship between X and Y is linear



The parametric approach outperforms the nonparametric approach

Case 2) True relationship between X and Y is strongly non-linear



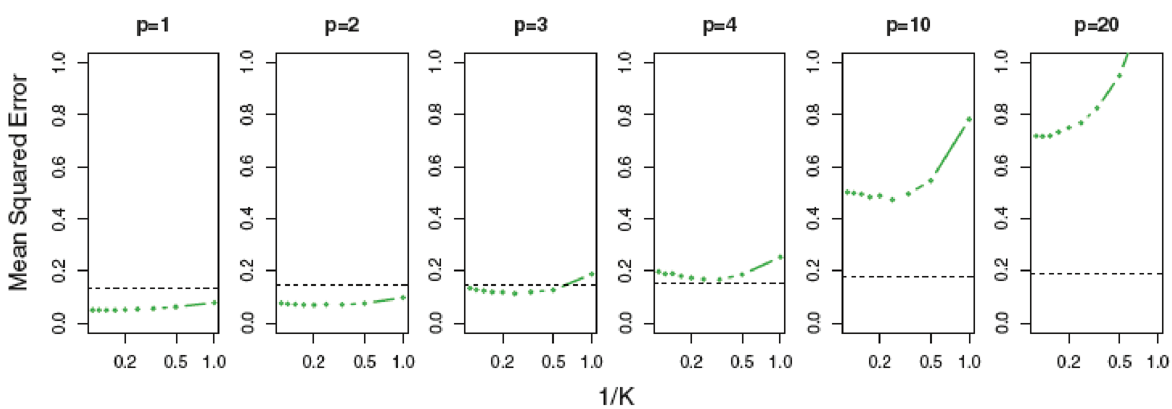
Instructor: Junho Song
junhosong@snu.ac.kr

The nonparametric approach outperforms the parametric approach

In general, a parametric model is superior to nonparametric model if the assumptions are valid.

```
install.packages('FNN') # To use 'knn.reg' in 'FNN' package
library('FNN')
library('ISLR')

Auto_s_hp = Auto[order(Auto$horsepower),] #Sort Auto dataset w.r.t.
   horsepower
knn.mpg.hp1 = knn.reg(train=Auto_s_hp$horsepower, test=NULL,
   y=Auto_s_hp$mpg, k=1, algorithm=c("kd_tree","cover_tree","brute"))
knn.mpg.hp9 = knn.reg(train=Auto_s_hp$horsepower, test=NULL,
   y=Auto_s_hp$mpg, k=9, algorithm=c("kd_tree","cover_tree","brute"))
knn.mpg.hp15 = knn.reg(train=Auto_s_hp$horsepower, test=NULL,
   y=Auto_s_hp$mpg, k=15, algorithm=c("kd_tree","cover_tree","brute"))

plot(horsepower,mpg)
lines(horsepower,knn.mpg.hp1$pred,type='l',col='blue')
lines(horsepower,knn.mpg.hp9$pred,type='l',col='red')
lines(horsepower,knn.mpg.hp15$pred,type='l',col='green')
```

(d) Curse of dimensionality (in KNN regression)



In high-dimensional problems, a given observation x_0 may have no nearby neighbors → "curse of dimensionality" (i.e. sparse coverage of the high-dimensional predictor space by data) → leading to a poor prediction of $f(x_0)$