IST 5535: Machine Learning Algorithms and Applications

Langtao Chen, Spring 2021

3. Linear Regression

Reading

- Book Chapter 3

Learning Objectives

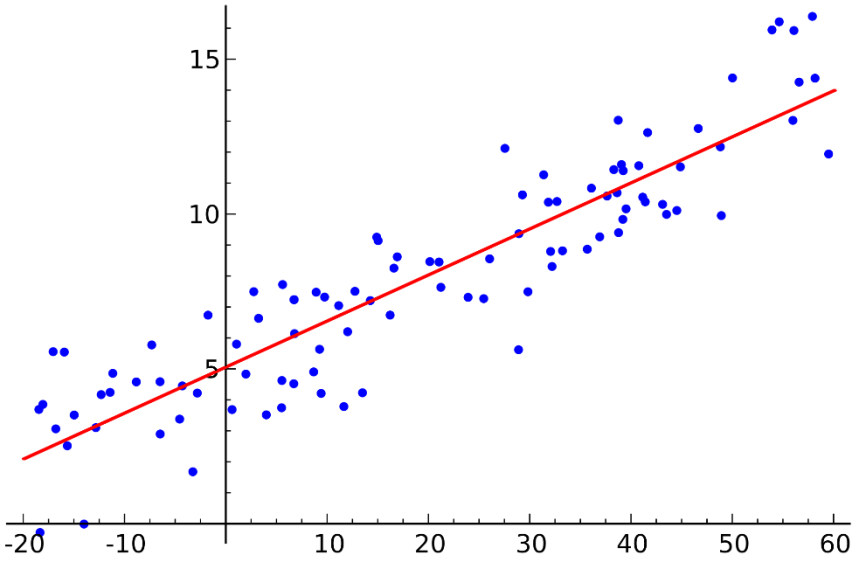
- Understand linear regression coefficient estimation and the ways of assessing the accuracy of coefficient estimates and the accuracy of the model.
- Understand methods dealing with qualitative predictors in linear regression.
- Understand interaction terms in linear regression.
- Understand non-linear relationship fit using polynomial regression.
- Understand potential problems of linear regression.
- Understand the comparison between linear regression and KNN regression.
- Be able to use R to conduct linear regression analysis and use diagnostic plots to check potential issues in linear regression.

Agenda

- Linear Regression
  - Estimating the Coefficients
  - Assessing the Accuracy of Coefficient Estimates
  - Assessing the Accuracy of the Model
- Other Considerations in Regression Model
  - Qualitative Predictors
  - Extensions of the Linear Model
  - Potential Problems
- Linear Regression vs. KNN

What is Regression?

- Regression is about estimating relationships between variables.
- Regression is a statistical technique that attempts to build a function of independent variables (regressors, input variables, or predictors) to predict or explain a dependent variable (response, or outcome).
- Regression intends to summarize observed data as simply and usefully as possible.



Major Objectives of Regression Analysis

- Explanatory modeling
  - The purpose is to explain or quantify the effect of independent variables on dependent variable
  - The classical statistical approach
  - Focus on unveiling the underlying relationship between variables
  - Use the entire dataset to fit the model with the data

- Predictive modeling
  - Predict the outcome value for new records, given value(s) of their input variable(s)
  - Focus on predictive performance rather than coefficients (beta)
  - Train the model on a training dataset and evaluate its performance on a test dataset

Linear Regression Model

- Linear regression model is a special case of the parametric model

$$y = X\beta + \varepsilon$$

where

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

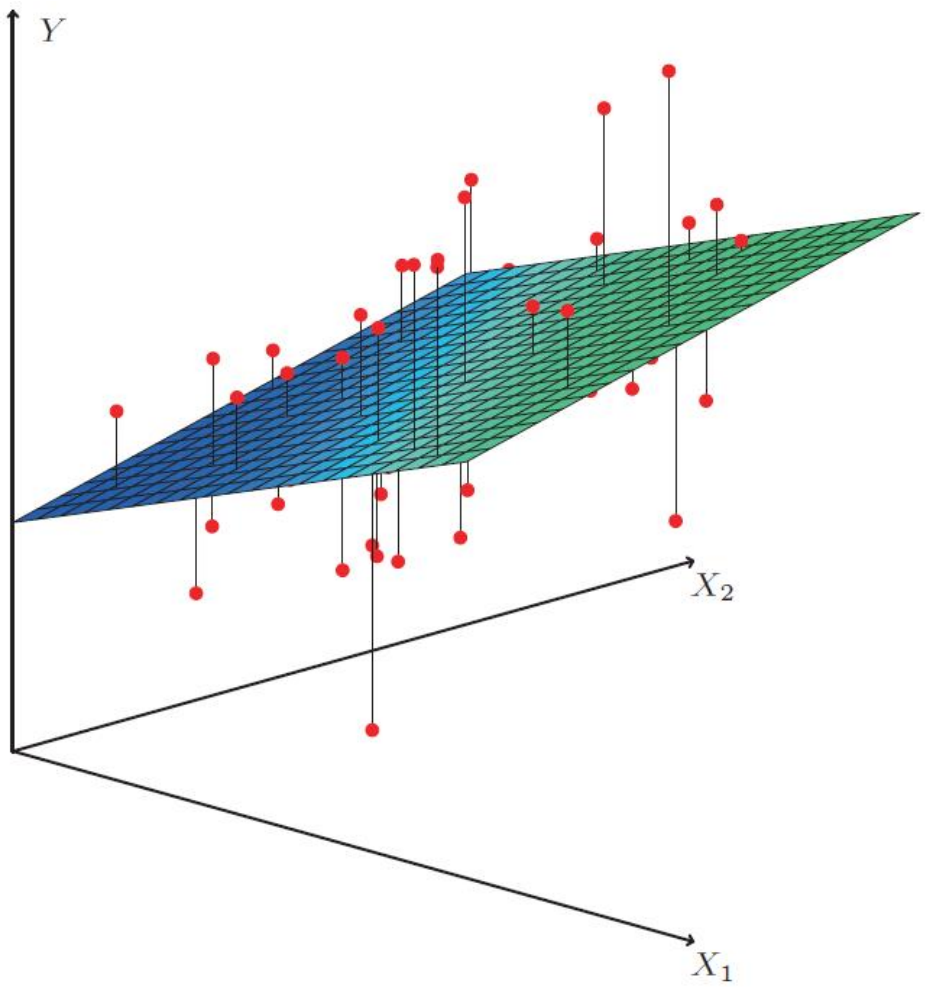
- Benefits
  - The simple linear functional form is easy to estimate
  - Interpretation is straightforward

Estimate Linear Regression Parameters

- Use ordinary least squares (OLS) to find $\hat{\beta}$ that minimize MSE

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{\beta}X\right)^2$$



Formally Define OLS: Optimization Problem

Set $y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$ $X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$ $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}$

Then $SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = (y - X\beta)^T (y - X\beta)$

To find the best model is to find the parameters $\hat{\beta}$ that minimize the SSE (error sum of squares).

Formal definition:

$$\hat{\beta} = arg \min_{\beta} [(y - X\beta)^T (y - X\beta)]$$

Ordinary Least Squares (OLS) Estimator

Target: $\hat{\beta} = arg\min_{\beta}[(y - X\beta)^T(y - X\beta)]$

According to optimization theory, the optimal parameter $\hat{\beta}$ satisfies the following conditions:

1. First order condition (F.O.C.):

$$\frac{\partial[(y - X\beta)^T(y - X\beta)]}{\partial\beta} = 0 \Leftrightarrow \frac{\partial[y^Ty - 2\beta^TX^Ty + \beta^TX^TX\beta]}{\partial\beta} = 0 \Leftrightarrow X^TX\hat{\beta} - X^Ty = 0$$

$$\Leftrightarrow X^TX\hat{\beta} = X^Ty$$

If X^TX is invertible, $\hat{\beta} = (X^TX)^{-1}X^Ty$.

2. Second order condition (S.O.C.): $\frac{\partial^2[(y - X\beta)^T(y - X\beta)]}{\partial\beta\partial\beta^T} = X^TX \geqslant 0$ (positive semi definite)

Thus, $\hat{\beta} = (X^TX)^{-1}X^Ty$ minimizes the SSE. This is called the OLS estimator (closed form solution).

Simple Linear Regression

$$y = \beta_0 + \beta_1 x$$

- One dependent variable (y): the one to predict or explain
- One independent variable (x): explanatory variable/predictor
- β_0: intercept
  - When x equals to zero, what is the value of y.
- β_1: slope
  - Increase x by one unit, how much would y change.

Multiple Linear Regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ \ldots\ldots + \beta_p x_p$$

- One dependent variable (y): the one to predict or explain
- Multiple independent variables (x_1, x_2, … x_n): explanatory
- β_0 : intercept
  - When all explanatory variables are zero, what is the value of y.
- β_i : slope (i >= 1)
  - Increase x_i by one unit, how much would y change after controlling for other factors.

Inference in Regression

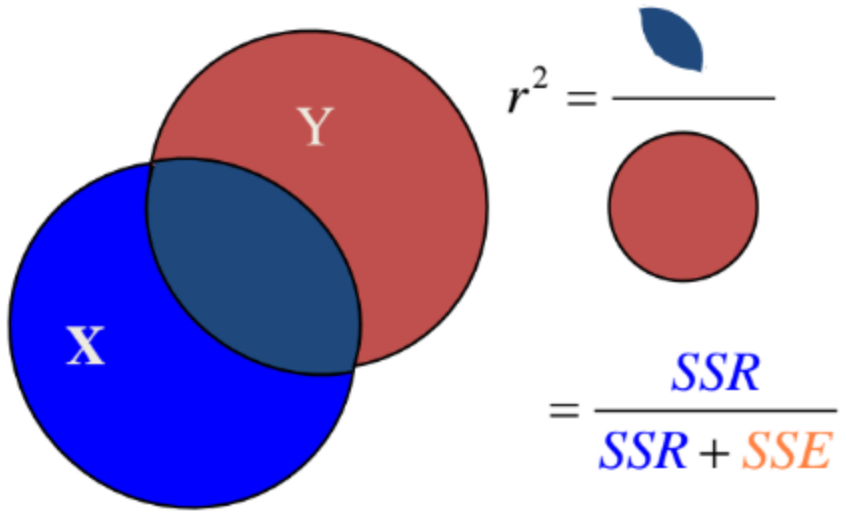
- How well does the regression model fit the data?
- What is the relationship between X and Y?
- What is the expected value of Y given an X value?

Measure of Fit: R^2

- The proportion of variation in Y that is explained by the independent variable X in the regression model.

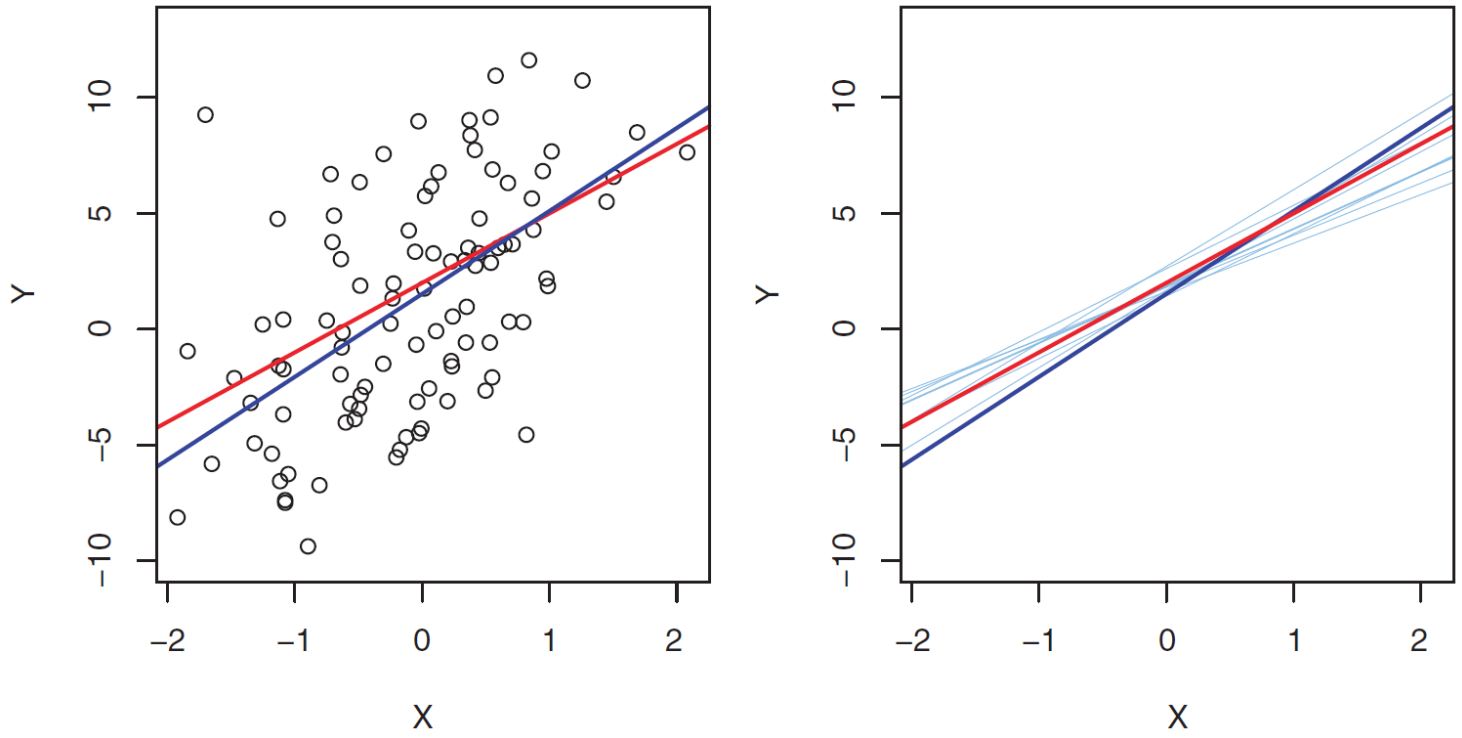
$$R^2 = 1 - \frac{SSE}{\sum(Y_i - \bar{Y})^2} = 1 - \frac{residual\ sum\ of\ squares}{total\ variance\ of\ Y}$$

- $R^2 \in [0, 1]$
  - $R^2 = 0$: X does not explain any variance of Y
  - $R^2 = 1$: X fully explains variance of Y (perfect fit)



Assessing the Accuracy of the Coefficient Estimates

- Different datasets result in slightly different OLS estimates.



- Red: population regression line (Y = 2 +3X) which is unknown in real dataset
- Dark Blue: OLS line estimated from observed data
- Light Blue: OLS line estimated from a random set of observations

Assessing the Accuracy of the Coefficient Estimates

- Population regression line (unknown)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ \ldots\ldots + \beta_p x_p$$

- OLS line

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x_1 + \widehat{\beta_2} x_2 + \ \ldots\ldots + \widehat{\beta_p} x_p$$

Statistical Inference is all about using an estimate from observed data to guess the true parameters.

We can calculate 95% confidence interval for β_j:

$$\left[\widehat{\beta_j} - 2 * SE\left(\widehat{\beta_j}\right), \widehat{\beta_j} + 2 * SE\left(\widehat{\beta_j}\right)\right]$$

The range contains the true value of the parameter with 95% probability.

Hypothesis Tests on Coefficients

- Null hypothesis

  H_0: There is no relationship between X_j and Y

- Alternative hypothesis

  H_a: There is some relationship between X_j and Y

Mathematically,

$$H_0: \beta_j = 0$$
$$H_a: \beta_j \neq 0$$

- In practice, we conduct a t-test to assess whether there is a relationship between X_j and Y:

$$t = \frac{\hat{\beta}_j - 0}{SE(\hat{\beta}_j)}$$

Example

```
> lm.fit=lm(sales ~ TV + radio + newspaper, data = Advertising)
> summary(lm.fit)

Call:
lm(formula = sales ~ TV + radio + newspaper, data = Advertising)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8277 -0.8908  0.2418  1.1893  2.8292

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.938889   0.311908   9.422   <2e-16 ***
TV           0.045765   0.001395  32.809   <2e-16 ***
radio        0.188530   0.008611  21.893   <2e-16 ***
newspaper   -0.001037   0.005871  -0.177     0.86
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared:  0.8972,  Adjusted R-squared:  0.8956
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

1. How well does the model fit the data?

```
> stargazer::stargazer(lm.fit, type = "text")
================================
                  Dependent variable:
                      sales
--------------------------------
TV                   0.046***
                    (0.001)
radio                0.189***
                    (0.009)
newspaper           -0.001
                    (0.006)
Constant             2.939***
                    (0.312)
--------------------------------
Observations          200
R2                    0.897
Adjusted R2           0.896
Residual Std. Error   1.686 (df = 196)
F Statistic           570.271*** (df = 3; 196)
================================
Note:           *p<0.1; **p<0.05; ***p<0.01
```

89.7% of the variance of sales can be explained by TV, radio, and newspaper advertising budgets.

2. What is the relationship between X and Y?

```
> stargazer::stargazer(lm.fit, type = "text")
================================
                  Dependent variable:
                  sales
--------------------------------
TV                0.046***
                  (0.001)
radio             0.189***
                  (0.009)
newspaper         -0.001
                  (0.006)
Constant          2.939***
                  (0.312)
--------------------------------
Observations      200
R2                0.897
Adjusted R2       0.896
Residual Std. Error  1.686 (df = 196)
F Statistic       570.271*** (df = 3; 196)
================================
Note:             *p<0.1; **p<0.05; ***p<0.01
```

- P-value <0.001
  X and Y are statistically significantly related at an alpha level of 0.001
- P-value <0.01
  X and Y are statistically significantly related at an alpha level of 0.01
- P-value <0.05
  X and Y are statistically significantly related at an alpha level of 0.05

TV budget and sales are statistically significantly related at an alpha level of 0.01, controlling for other factors.

Radio budget and sales are statistically significantly related at an alpha level of 0.01, controlling for other factors.

Newspaper budget and sales are NOT statistically significantly related, controlling for other factors.

3. What is the expected value of Y given an X value?

```
> stargazer::stargazer(lm.fit, type = "text")
================================
                  Dependent variable:
                      sales
--------------------------------
TV                    0.046***
                      (0.001)
radio                 0.189***
                      (0.009)
newspaper             -0.001
                      (0.006)
Constant              2.939***
                      (0.312)
--------------------------------
Observations            200
R2                      0.897
Adjusted R2             0.896
Residual Std. Error     1.686 (df = 196)
F Statistic         570.271*** (df = 3; 196)
================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

$$\widehat{sales} = 2.939 + 0.046 * TV + 0.189 * radio - 0.001 * newspaper$$

Predict sales when allocating all \$300k budgets to TV

$$\widehat{sales} = 2.939 + 0.046 * 300 + 0.189 * 0 - 0.001 * 0$$
$$= 16.739 \text{ (thousand units)}$$

Agenda

- Linear Regression
  - Estimating the Coefficients
  - Assessing the Accuracy of Coefficient Estimates
  - Assessing the Accuracy of the Model
- Other Considerations in Regression Model
  - Qualitative Predictors
  - Extensions of the Linear Model
  - Potential Problems
- Linear Regression vs. KNN

Qualitative Predictors (a.k.a. Factors)

- From the Credit dataset, we want to investigate differences in credit card balance between males and females.

```
> str(Credit)
'data.frame': 400 obs. of  12 variables:
 $ ID        : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Income    : num  14.9 106 104.6 148.9 55.9 ...
 $ Limit     : int  3606 6645 7075 9504 4897 8047 3388 7114 3300 ...
 $ Rating    : int  283 483 514 681 357 569 259 512 266 491 ...
 $ Cards     : int  2 3 4 3 2 4 2 2 5 3 ...
 $ Age       : int  34 82 71 36 68 77 37 87 66 41 ...
 $ Education: int  11 15 11 11 16 10 12 9 13 19 ...
 $ Gender    : Factor w/ 2 levels " Male","Female": 1 2 1 2 1 1 2  ...
 $ Student   : Factor w/ 2 levels "No","Yes": 1 2 1 1 1 1 1 1 1 2 ...
 $ Married   : Factor w/ 2 levels "No","Yes": 2 2 1 1 2 1 1 1 1 2 ...
 $ Ethnicity: Factor w/ 3 levels "African American",..: 3 2 2 2 3 ...
 $ Balance   : int  333 903 580 964 331 1151 203 872 279 1350 ...
```

Code Factors as Dummy Variables or Indicator Variables

- Qualitative Predictors with Two Levels
  - For example gender

$$female_i = \begin{cases} 1 & \text{if the } i\text{th person is female} \\ 0 & \text{if the } i\text{th person is male} \end{cases}$$

- Qualitative Predictors with More than Two Levels
  - For example ethnicity
  - Create multiple dummy variables

$$ethnicity\_asian_i = \begin{cases} 1 & \text{if the } i\text{th person is Asian} \\ 0 & \text{if the } i\text{th person is not Asian} \end{cases}$$

$$ethnicity\_caucasian_i = \begin{cases} 1 & \text{if the } i\text{th person is Caucasian} \\ 0 & \text{if the } i\text{th person is not Caucasian} \end{cases}$$

Extension of the Linear Model

- Two basic assumptions of linear model
  - Additive assumption: the effect of X_j on Y is independent of other predictors.
  - Linear assumption: the change in Y due to a one-unit change of X_j is constant, regardless of the value of X_j.

Removing the Additive Assumption

- Interaction effect: the effect of X_j on Y is dependent of another predictor X_k
- For example, perhaps spending $50,000 on television advertising and $50,000 on radio advertising results in more sales than allocating $100,000 to either television or radio individually.
- In marketing, this is called a ***synergy*** effect.

Interaction in Advertising

$$sales = \beta_0 + \beta_1 * TV + \beta_2 * Radio + \beta_3 * TV * Radio + \epsilon$$

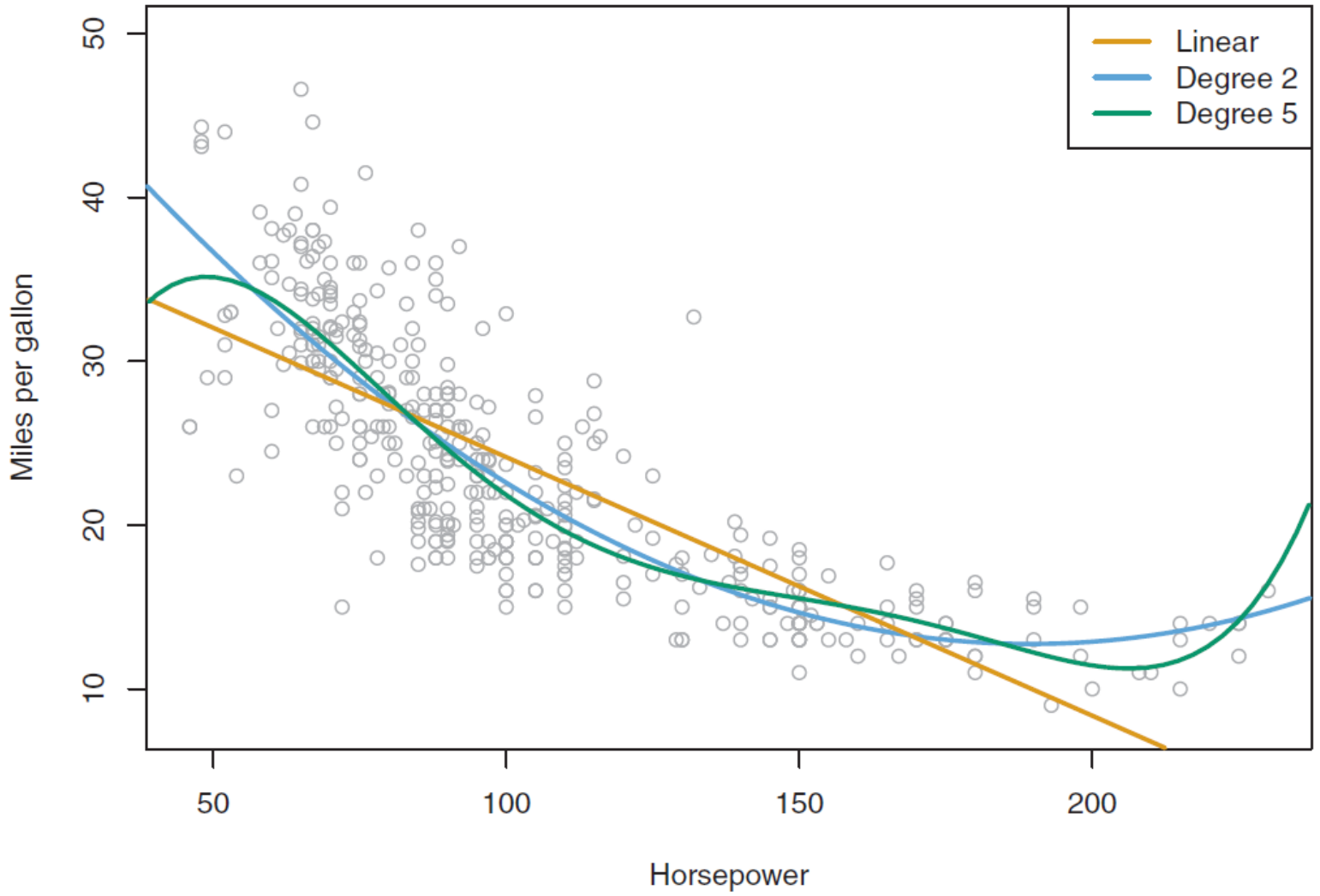
	Coefficient	Std. error	t-statistic	p-value
Intercept	6.7502	0.248	27.23	< 0.0001
TV	0.0191	0.002	12.70	< 0.0001
radio	0.0289	0.009	3.24	0.0014
TV×radio	0.0011	0.000	20.73	< 0.0001

$$sales = \beta_0 + (\beta_1 + \beta_3 * Radio) * TV + \beta_2 * Radio + \epsilon$$

The effect of TV on sales depends on radio advertising:

$$\tilde{\beta}_1 = \beta_1 + \beta_3 * Radio$$

Non-Linear Relationships



Use Linear Model to Fit Nonlinear Relationship

- Dependent variable y is modeled as an *h-th* degree polynomial of x:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots\ldots + \beta_h x^h + \varepsilon$$

- Relationships when the degree h is low:
  - h=2: quadratic
  - h=3: cubic
  - h=4: quartic

- Polynomial regression is a linear model, since the outcome y is a linear combination of coefficients $\beta_i (i = 1, 2, \ldots h)$

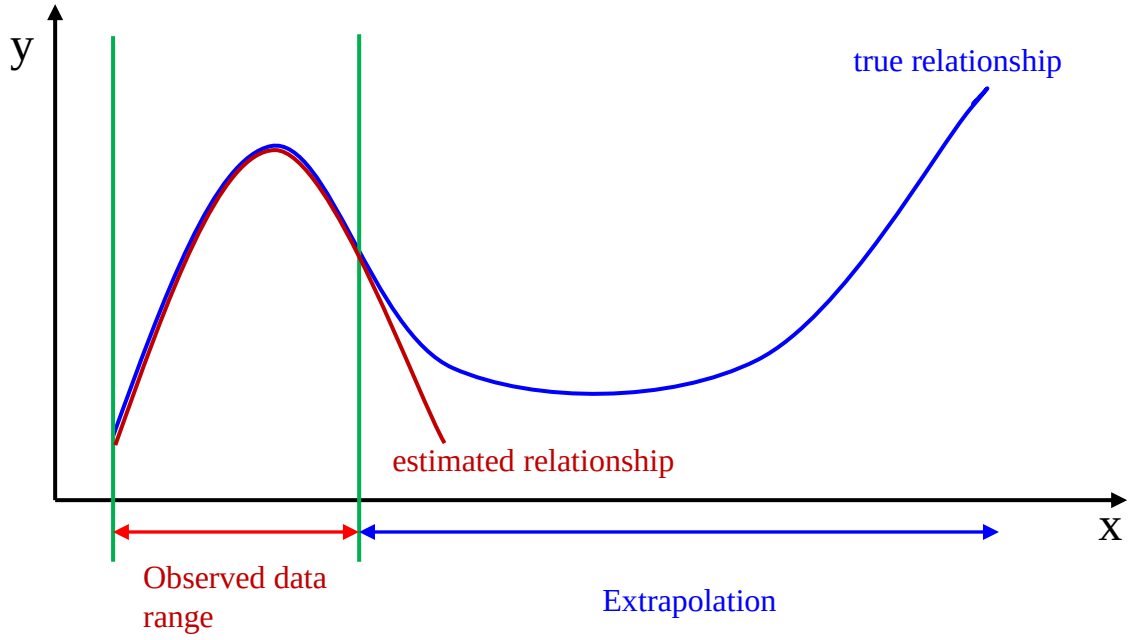
The Hierarchy Principle:

If the polynomial regression model contains x^h and its coefficient is significant, then the model should also include all lower-degree terms $x^j (j < h)$, no matter those x^j are significant or not.

Be Cautious of the Overfitting Issue

- Polynomial regression may be misleading if you don't have a large dataset.

- Do NOT extrapolate beyond your observed data range.



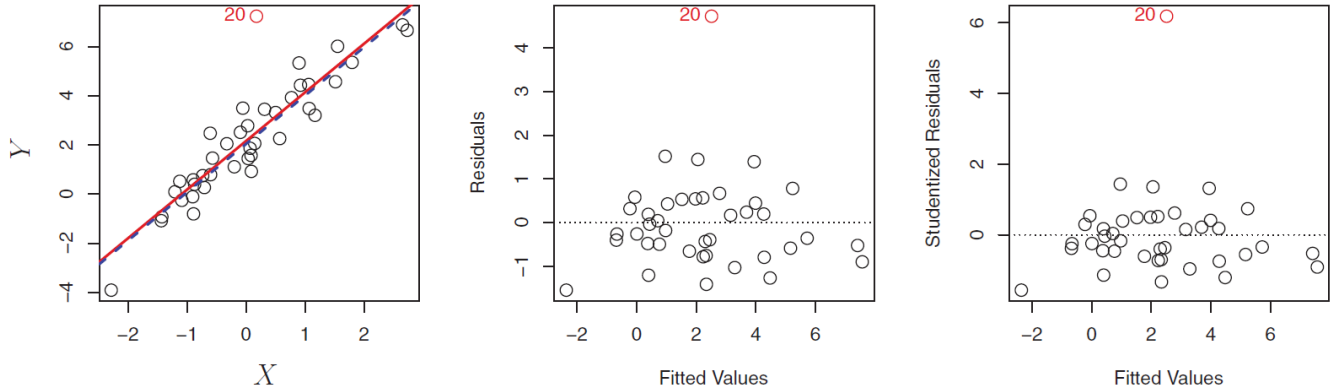
Outside the observed data range, inference or prediction is not reliable.

Potential Fit Problems

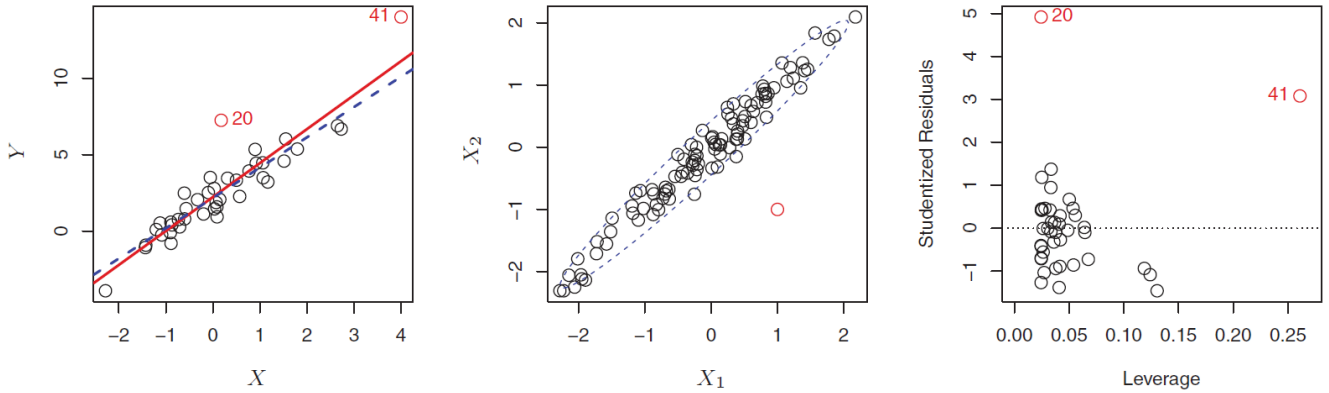
- When we fit a linear regression model to a particular data set, many problems may occur. Most common among these are:
  - 1. *Non-linearity of the response-predictor relationships*
  - 2. *Correlation of error terms*
  - 3. *Non-constant variance of error terms (heteroscedasticity)*
  - 4. *Outliers*
  - 5. *High-leverage points*
  - 6. *Collinearity*

Influential Points

- Outliers: data points with unusually large/small response values



- High leverage points: data points with unusually large/small independent values



$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n}(x_{i'} - \bar{x})^2}$$

Agenda

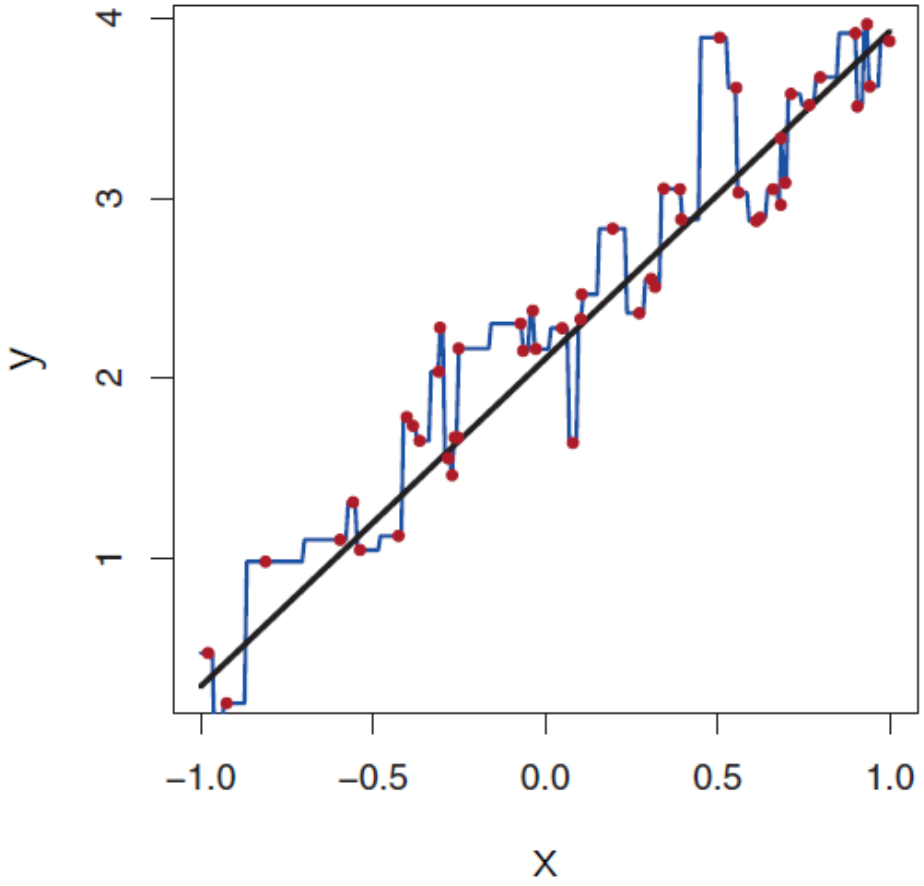
- Linear Regression
  - Estimating the Coefficients
  - Assessing the Accuracy of Coefficient Estimates
  - Assessing the Accuracy of the Model
- Other Considerations in Regression Model
  - Qualitative Predictors
  - Extensions of the Linear Model
  - Potential Problems
- Linear Regression vs. KNN

KNN Regression

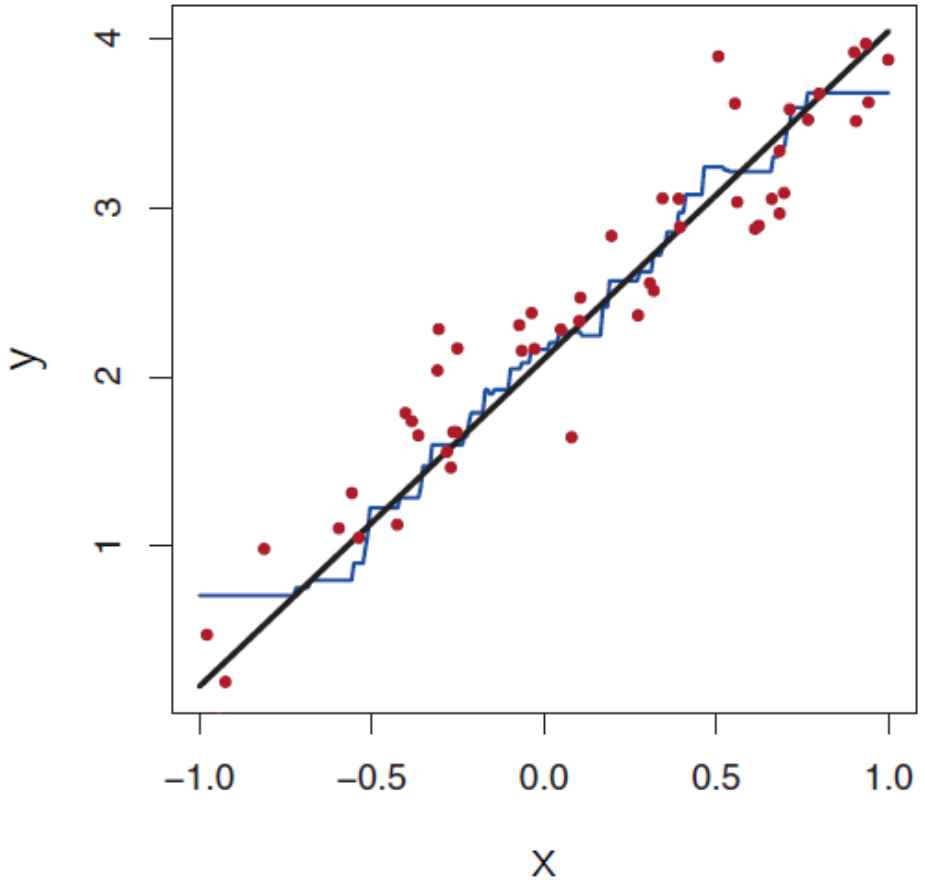
- KNN regression works similar to KNN classification
  - Step 1: Find *K* nearest neighbors;
  - Step 2: Predict the response as the average of *K* neighbors:

$$\hat{f}\,(x_0) = \frac{1}{K} \sum_{x_0 \in N_0} y_i$$

Larger K results in smoother fit



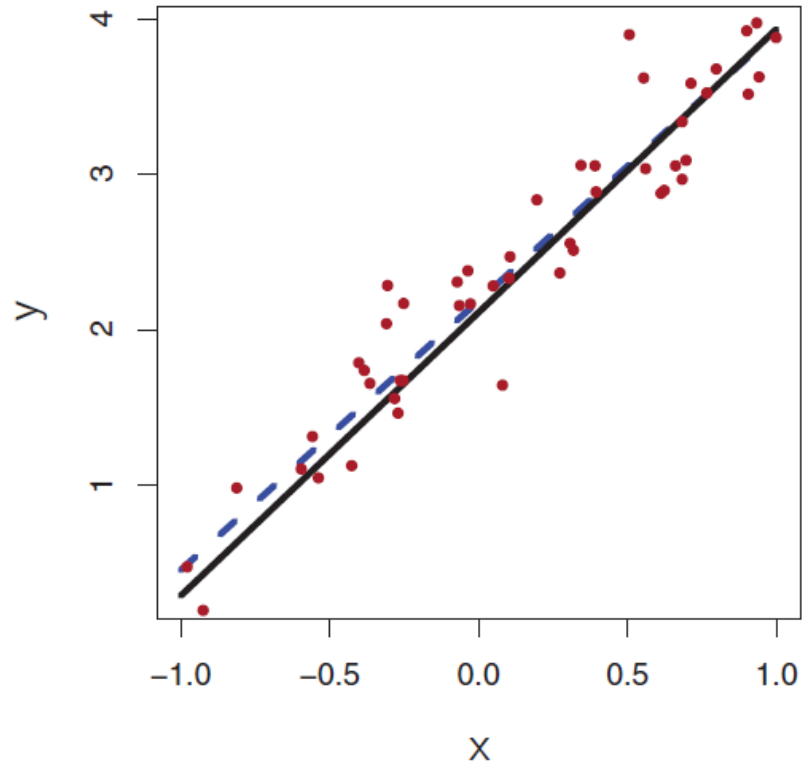
K=1

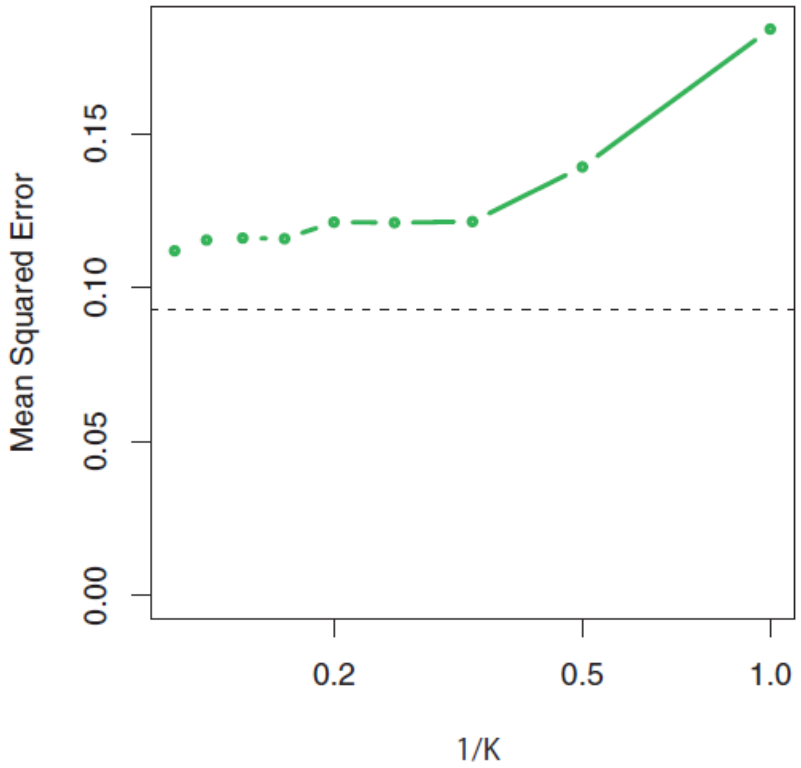


K=9

True Linear Relation, One Dimension (p=1)

- When K is large, KNN performs only a little worse than linear regression;
- When K is small, KNN performs far worse.



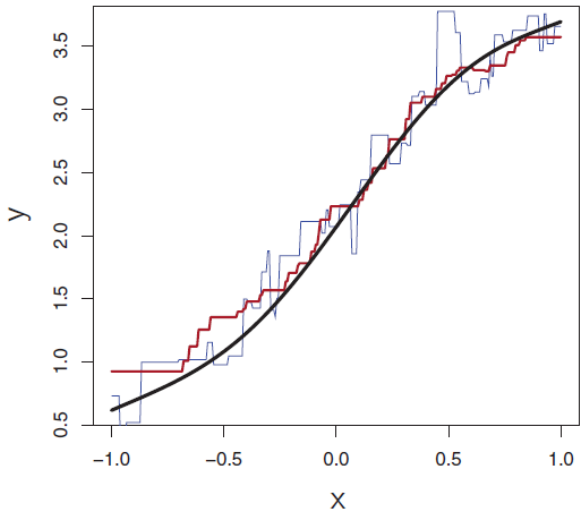


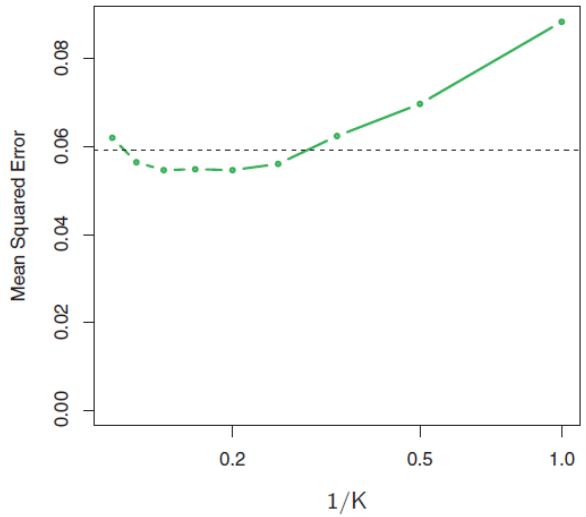
KNN regression

Linear regression

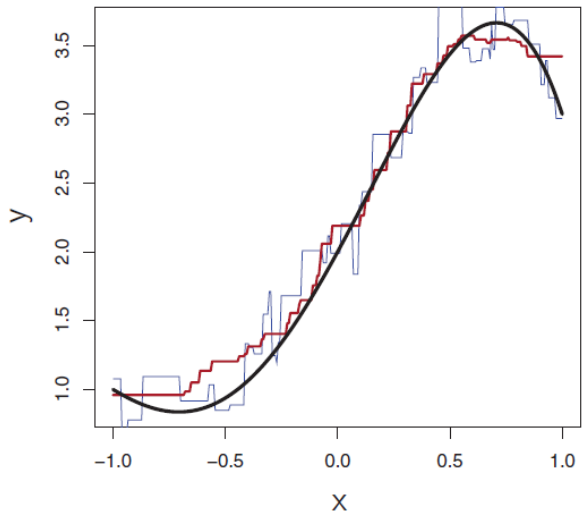
True Nonlinear Relation, One Dimension (p=1)

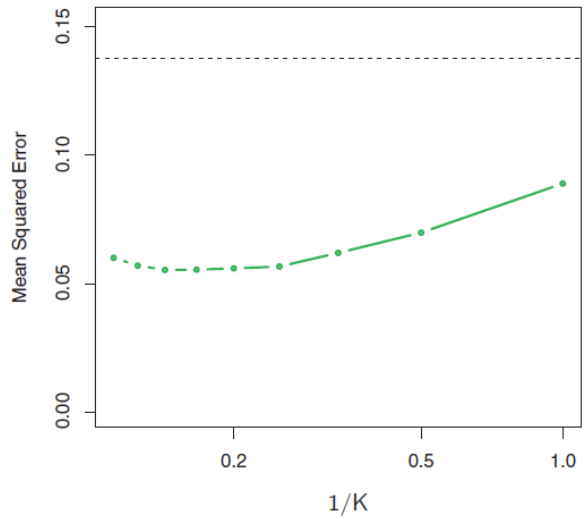
- Slightly non-linear relationship





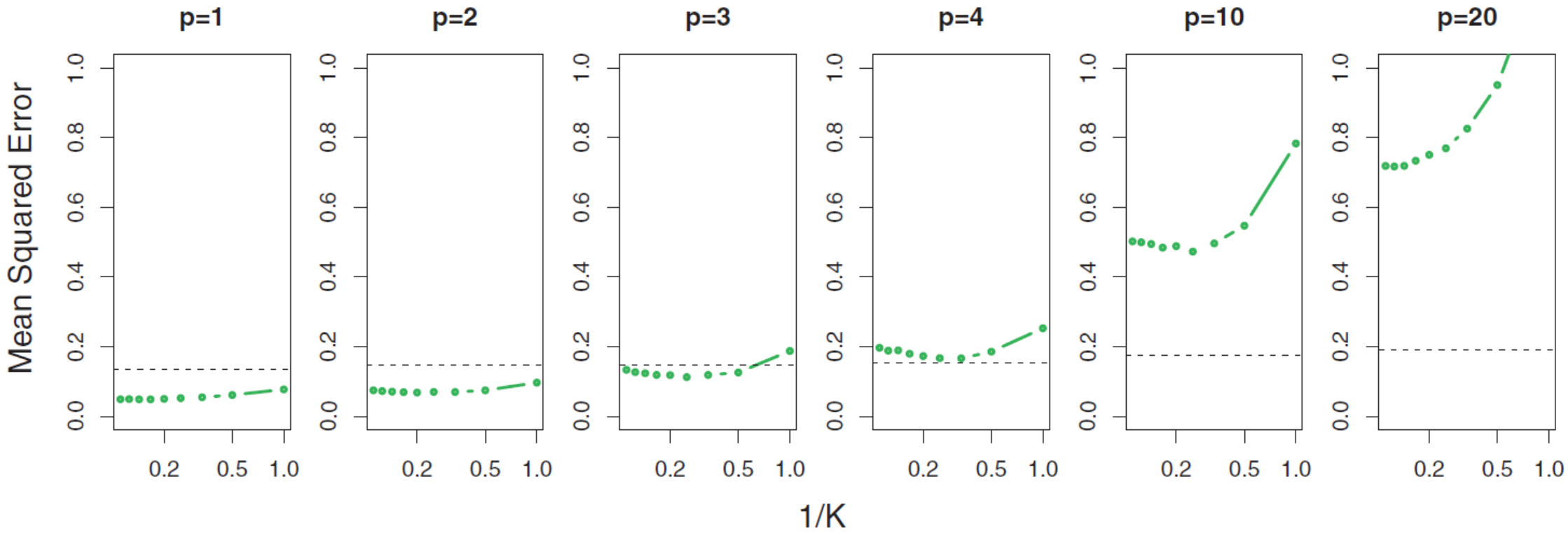
- Strongly non-linear relationship





True Nonlinear Relation, Different Dimensions

- KNN suffers from the curse of dimensionality.



Review: Learning Objectives

- Understand linear regression coefficient estimation and the ways of assessing the accuracy of coefficient estimates and the accuracy of the model.
- Understand methods dealing with qualitative predictors in linear regression.
- Understand interaction terms in linear regression.
- Understand non-linear relationship fit using polynomial regression.
- Understand potential problems of linear regression.
- Understand the comparison between linear regression and KNN regression.
- Be able to use R to conduct linear regression analysis and use diagnostic plots to check potential issues in linear regression.

Q & A