LINEAR REGRESSION

Chapter 03

Outline

- The Linear Regression Model
  - Least Squares Model Fitting
  - Measures of Fit
  - Inference in Regression
- Other Considerations in Regression Model
  - Qualitative Predictors
  - Interaction Terms
- Potential Fit Problems
- Linear vs. KNN Regression

Outline

- The Linear Regression Model
  - Least Squares Model Fitting
  - Measures of Fit
  - Inference in Regression
- Other Considerations in Regression Model
  - Qualitative Predictors
  - Interaction Terms
- Potential Fit Problems
- Linear vs. KNN Regression

The (multiple) Linear Regression Model

$$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi} + \epsilon$$

- The parameters in the linear regression model are very easy to interpret.
- β_0 is the intercept (i.e. the average value for Y if all the X's are zero), β_j is the slope for the j^{th} variable X_j
- β_j is the average increase in Y when X_j is increased by one and **all other X's are held constant.**

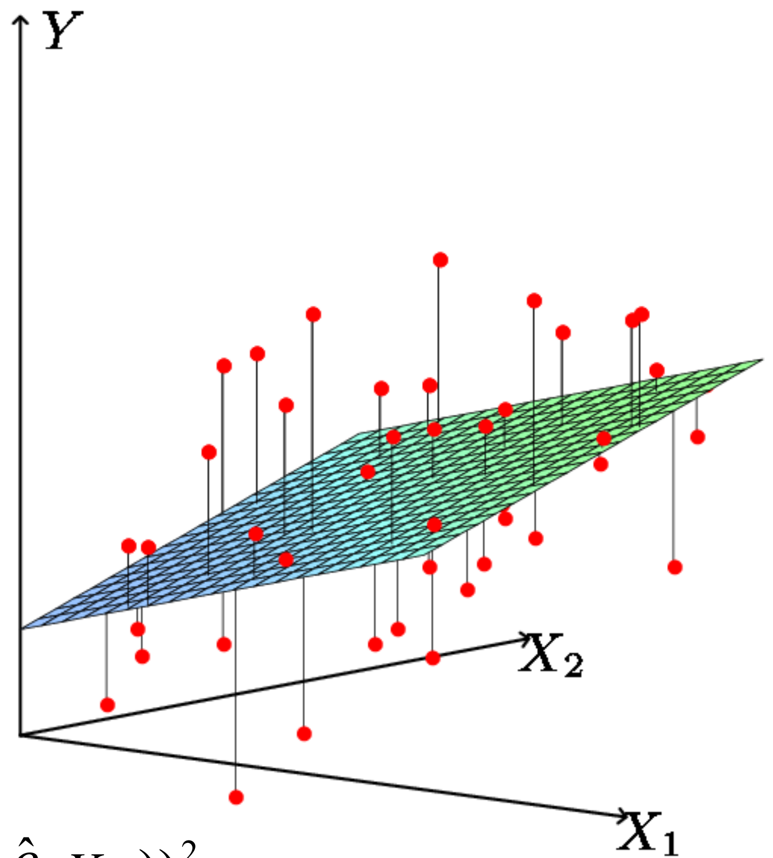
Least Squares Fit

- Estimate the parameters using least squares

- The best coeff's are the ones which minimize the cost

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

$$= \frac{1}{n}\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \ldots + \hat{\beta}_p X_{pi}))^2$$



Concept Check:
What is the difference between RSS and MSE?

Relationship between population and least squares fit

Population $$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \acute{U}$$

Least Squares fit $$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \ldots + \hat{\beta}_p X_{pi}$$

- Would like to know β_0 through β_p: the population line. Instead we know $\hat{\beta}_0$ through $\hat{\beta}_p$: the least squares line.
- Use $\hat{\beta}_0$ through $\hat{\beta}_p$ as guesses for β_0 through β_p and $\hat{y}_i$ as a guess for y.

Least Squares Pseudocode Exercise

- Write pseudocode for a primitive method for determining the least-squares model fit in 1-variable linear regression (to find β_0 & β_1)
  - Your observations are stored in matrix X. For each observation, assume you are given x_1 and the corresponding y.
  - Hint: If you want to do gradient descent, you could compute a "local gradient" near a value of β_i by computing the RSS change occurring from an epsilon increase of the coefficient:
    - RSS when using ($\beta_i+\varepsilon$) minus RSS when using ($\beta_i-\varepsilon$)
  - Think: how would you use these local gradients to search for a best set of beta values?
- How would you extend your idea to a general multiple linear regression model fitting algorithm?

Least Squares Python Exercise

- Write python code for a primitive method for determining the least-squares model fit in 1-variable linear regression (to find β_0 & β_1)
  - Your observations are stored in matrix X. For each observation, assume you are given x_1 and the corresponding y.
- Your portion of the code needs to compute a "local gradient" near a value of β_i by computing the RSS change occurring from an epsilon increase of the coefficient (for each coefficient):
  - RSS(f(X at $\beta_0+\varepsilon, \beta_1$))-RSS(f(X at $\beta_0-\varepsilon, \beta_1$))
  - RSS(f(X at $\beta_0, \beta_1+\varepsilon$))-RSS(f(X at $\beta_0, \beta_1-\varepsilon$))

Evaluation Criteria Worksheet

There are a number of evaluation criteria for linear regression models. Fill out the first side of the handout per the instructions

RSS	*p*-value
MSE	R^2
TSS	Correlation(X,Y)
Var & SE	F-statistic
RSE	Leverage statistic
t-statistic	VIF

We will discuss a subset of these in class

Measure of Lack of Fit: Residual Standard Error (RSE)

- RSE is an estimate of the standard deviation of the irreducible error ε.
- Roughly the average amount that the response will deviate from the true regression line (because of ε)

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

RSE is sensitive to the Y scale of the data since it is measured in units of y.

Measures of Fit: R^2

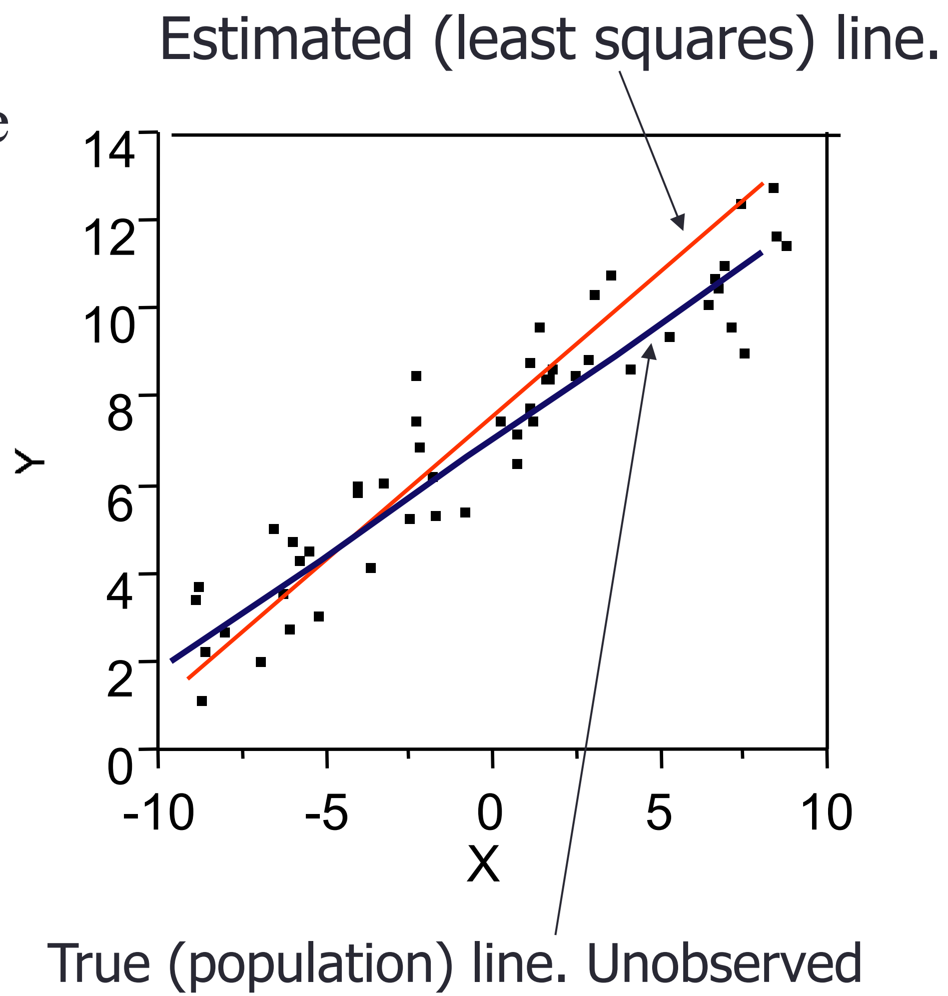
- Some of the variation in Y can be explained by variation in the X's and some cannot.
- R^2 is a proportion of the variance and is scale invariant
- R^2 tells you the fraction of variance that can be explained by X.

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \approx 1 - \frac{\text{Ending Variance}}{\text{Starting Variance}}$$

R^2 is always between 0 and 1. Zero means no variance of the response (Y) has been explained by the model. One means all the variance in the response Y has been explained (perfect fit to the data).

Prediction & Inference in Regression

- The regression line from the sample is not the regression line from the population.
- What we want to do:
  - Guess what value Y would take for a given X value
  - Assess how well the line describes the plot.
  - Guess the slope of the population line.



Feature (Predictor) Relevance

- Can we be sure that at least one of our X variables is a useful predictor? [i.e. not the case that $\beta_1 = \beta_2 = \cdots = \beta_p = 0$]
- Do all the predictors help to explain Y, or are only a subset useful?
  - In other words, is β_j=0 or not? We can use a hypothesis test to answer this question.
  - Feature Selection: If we can't be sure that β_j≠0 then there is no point in using X_j as one of our predictors.

Evaluating the regression model (1/2)

- Test for:
  - H_0: *all* slopes = 0 $(\beta_1 = \beta_2 = \cdots = \beta_p = 0)$,
  - H_a: at least one slope $\neq 0$
  - *p* predictors (features) and *n* observations
  - Compute the F statistic

$$F = \frac{\left(\frac{(TSS - RSS)}{p}\right)}{\left(\frac{RSS}{(n-p-1)}\right)}$$

When F is close to 1 there is no relationship between the response and the predictors
When F > 1, we can consider rejecting H_0
The amount above 1 required depends on *n*.
The larger *n* is, the less F has to be to reject H_0
Note: $p<n$ for this to be useful

Evaluating the regression model (2/2)

- Test for:
  - H_0: all slopes = 0 $(\beta_1 = \beta_2 = \cdots = \beta_p = 0)$,
  - H_a: at least one slope $\neq 0$

$$F = \frac{\frac{(TSS - RSS)}{p}}{\frac{RSS}{(n-p-1)}}$$

Answer comes from the F test in the **ANalysis Of VAriance** (ANOVA) table.
The ANOVA table has many pieces of information. What we care about is the F-Ratio and the corresponding *p*-value.

ANOVA Table

Source	df	SS	MS	F	p-value
Explained	2	4860.2347	2430.1174	859.6177	0.0000
Unexplained	197	556.9140	2.8270		

Given a passing F-test, Is $\beta_j \neq 0$? is X_j an important variable?

- We use a hypothesis test to answer this question
- H_0: $\beta_j = 0$ vs H_a: $\beta_j \neq 0$
- Calculate $\boxed{t = \hat{\beta}_j \big/ SE(\hat{\beta}_j)}$ ← Number of standard deviations away from zero.
- If t is large (equivalently p-value is small) we can be sure that $\beta_j \neq 0$ and that there is a relationship

Regression coefficients

	Coefficient	Std Err	t-value	p-value
Constant	7.0326	0.4578	15.3603	0.0000
TV	0.0475	0.0027	17.6676	0.0000

$\hat{\beta}_1$ $SE(\hat{\beta}_1)$ P-value

$\hat{\beta}_1$ is 17.67 SE's from 0

Testing Individual Variables & Conditional Relationships

Example: Is there a (statistically detectable) linear relationship between Newspapers and Sales given all the other variables have been accounted for? What about if Newspaper is the only available media?

Regression coefficients

	Coefficient	Std Err	t-value	p-value
Constant	2.9389	0.3119	9.4223	0.0000
TV	0.0458	0.0014	32.8086	0.0000
Radio	0.1885	0.0086	21.8935	0.0000
Newspaper	-0.0010	0.0059	-0.1767	0.8599

big p-value: NO

Regression coefficients

	Coefficient	Std Err	t-value	p-value
Constant	12.3514	0.6214	19.8761	0.0000
Newspaper	0.0547	0.0166	3.2996	0.0011

Small p-value in simple regression

Interpretation: Newspaper doesn't add much given that TV and Radio are used. Decision: If we can use TV & Radio, we should, but if they are not available, Newspaper still affects sales.

Outline

- The Linear Regression Model
  - Least Squares Model Fitting
  - Measures of Fit
  - Inference in Regression
- Other Considerations in Regression Model
  - Qualitative Predictors
  - Interaction Terms
- Potential Fit Problems
- Linear vs. KNN Regression

Two-way Qualitative Predictors

- Suppose you have a "gender" feature. How do you code "male" and "female" (category listings) into a regression equation?

- Option 1:
  Code them as indicator variables ("dummy" variables)
  - For example we can "code" Males=0 and Females= 1.

- Option 2:
  Code them as +1/-1 variables For example we can "code" Males= -1 and Females= 1.

Two-way Qualitative: Zero-One Coding

- Suppose we want to include income and gender to determine bank balance.
- Two genders (male and female). Let

$$\boxed{\text{Gender}_i = \begin{cases} 0 \text{ if male} \\ 1 \text{ if female} \end{cases}}$$

- then the regression equation is

$$\boxed{Y_i \approx \beta_0 + \beta_1 \text{Income}_i + \beta_2 Gender_i = \begin{cases} \beta_0 + \beta_1 Income_i & \text{if male} \\ \beta_0 + \beta_1 Income_i + \ \beta_2 & \text{if female} \end{cases}}$$

- Interpretation of β_2: The average extra balance each month that females have for given income level. Males

Regression coefficients

	Coefficient	Std Err	t-value	p-value
Constant	233.7663	39.5322	5.9133	0.0000
Income	0.0061	0.0006	10.4372	0.0000
Gender_Female	24.3108	40.8470	0.5952	0.5521

Two-way Qualitative: Other Coding Schemes

- There are different ways to code categorical variables.
- Two genders (male and female). Let

$$\boxed{Gender_{\mathrm{i}} = \begin{cases} -1 & \text{if male} \\ 1 & \text{if female} \end{cases}}$$

- then the regression equation is

$$\boxed{Y_i \approx \beta_0 + \beta_1 \mathrm{Income}_i + \beta_2 Gender_i = \begin{cases} \beta_0 + \beta_1 Income_i - \beta_2, & \text{if male} \\ \beta_0 + \beta_1 Income_i + \beta_2, & \text{if female} \end{cases}}$$

- Interpretation of β_2: The average amount that females are above the average, for any given income level. β_2 is also the average amount that males are below the average, for any given income level.

Multi-way Qualitative: Other Coding Schemes

- How would you code if there were more than 2 classes of a categorical variable
  - Example: color = {Red, Green, or Blue}
- Design a coding scheme and then explain how to interpret the resulting coefficients of your coding variables

Other Issues Discussed

- Interaction terms
- Non-linear effects
- Multicollinearity
- Model Selection

Interaction

- The effect on Y of increasing X_1 depends on another data feature (e.g. X_2)

- Example
  - The effect on Salary (Y) when increasing Position (X_1) also depends on gender (X_2)
  - Maybe as they get promoted, Male salaries go up faster (or slower) than Females.

- Advertising example:
  - TV and radio advertising both increase sales.
  - Perhaps due to synergy, spending money on both of them may increase sales more than spending the same amount on one alone?

Interaction in advertising

$$\boxed{Sales = \beta_0 + \beta_1 \times TV + \beta_2 \times Radio + \beta_3 \times TV \times Radio}$$

Interaction Term
TV & Radio together

$$\boxed{Sales = \beta_0 + (\beta_1 + \beta_3 \times Radio) \times TV + \beta_2 \times Radio}$$

- Spending \$1 extra on TV increases average sales by 0.0191 + 0.0011×Radio

$$\boxed{Sales = \beta_0 + (\beta_2 + \beta_3 \times TV) \times Radio + \beta_2 \times TV}$$

- Spending \$1 extra on Radio increases average sales by 0.0289 + 0.0011×TV

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	6.7502202	0.247871	27.23	<.0001*
TV	0.0191011	0.001504	12.70	<.0001*
Radio	0.0288603	0.008905	3.24	0.0014*
TV*Radio	0.0010865	5.242e-5	20.73	<.0001*

Should we consider interaction effects?

- Example: Relationship between job position and salary for men and women.
- Because we used a +1 / -1 dummy variable (gender), and did not include interaction terms, our model has forced the line for men and the line for women to be parallel.
- Parallel lines suggest that promotions have the same salary benefit for men as for women (even if that is not true in reality).
- Non-parallel line would suggest promotions affect men's and women's salaries differently

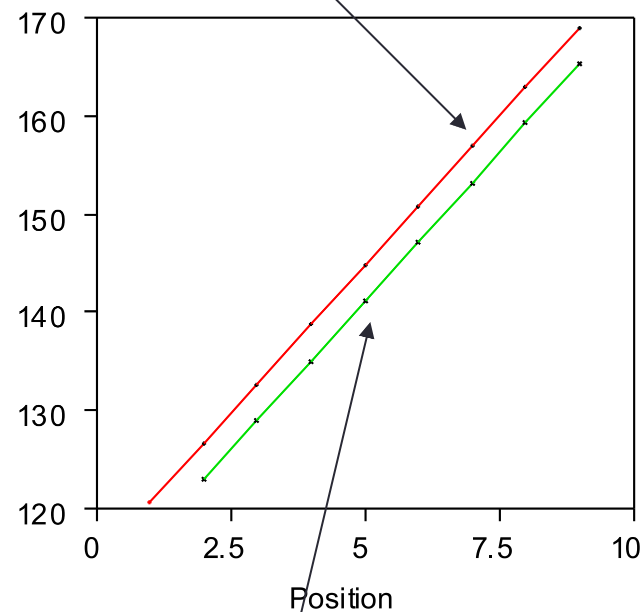
Parallel Regression Lines

Expanded Estimates

Nominal factors expanded to all levels

Term	Estimate	Std Error	t Ratio	Prob>|t
Intercept	112.77039	1.454773	77.52	<.0001
Gender[female]	1.8600957	0.527424	3.53	0.0005
Gender[male]	-1.860096	0.527424	-3.53	0.0005
Position	6.0553559	0.280318	21.60	<.0001

Line for women



Line for men

Regression equation

female: salary = 112.77+1.86 + 6.05× position

males: salary = 112.77-1.86 + 6.05 × position

Different intercepts

Same slopes

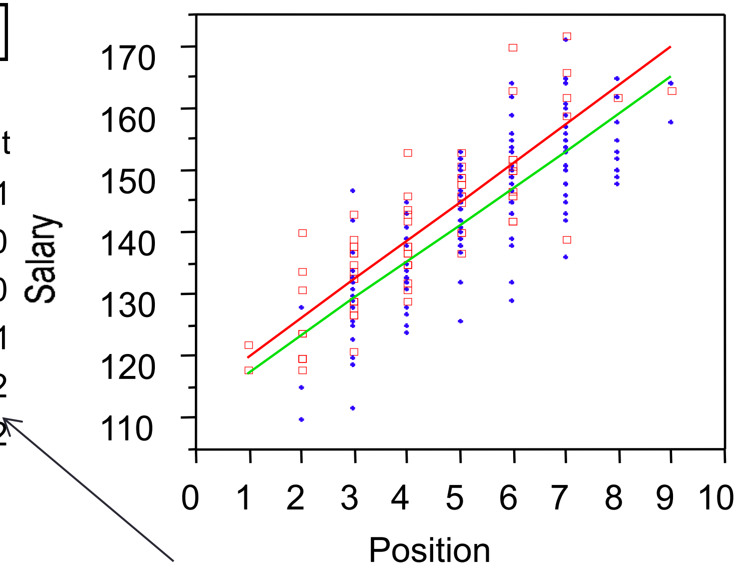
Parallel lines have the same slope. Dummy variables give lines different intercepts, but their slopes are still the same.

Should the Lines be Parallel?

Expanded Estimates

Nominal factors expanded to all levels

Term	Estimate	Std Error	t Ratio	Prob>|t
Intercept	112.63081	1.484825	75.85	<.0001
Gender[female]	1.1792165	1.484825	0.79	0.4280
Gender[male]	-1.179216	1.484825	-0.79	0.4280
Position	6.1021378	0.296554	20.58	<.0001
Gender[female]*Position	0.1455111	0.296554	0.49	0.6242
Gender[male]*Position	-0.145511	0.296554	-0.49	0.6242



Interaction between gender and position

Interaction is not significant

Procedure: Add interaction terms. Check for significance of coefficients.
Significant coeffs in this example are Intercept and Position.
Since gender-position interactions are not significant, no reason to reject parallel lines as a reasonable assumption
Interpretation: income increase due to promotions does not depend on gender

Outline

- The Linear Regression Model
  - Least Squares Model Fitting
  - Measures of Fit
  - Inference in Regression
- Other Considerations in Regression Model
  - Qualitative Predictors
  - Interaction Terms
- Potential Fit Problems
- Linear vs. KNN Regression

Potential Fit Problems Worksheet

There are a number of possible problems that one may encounter when fitting the linear regression model. Fill out the second side of the handout per the instructions

1. Non-linearity of the data
2. Dependence of the error terms
3. Non-constant variance of error terms
4. Outliers
5. High leverage points
6. Collinearity

See Section 3.3.3 for more details.

Outline

- The Linear Regression Model
  - Least Squares Model Fitting
  - Measures of Fit
  - Inference in Regression
- Other Considerations in Regression Model
  - Qualitative Predictors
  - Interaction Terms
- Potential Fit Problems
- Linear vs. KNN Regression

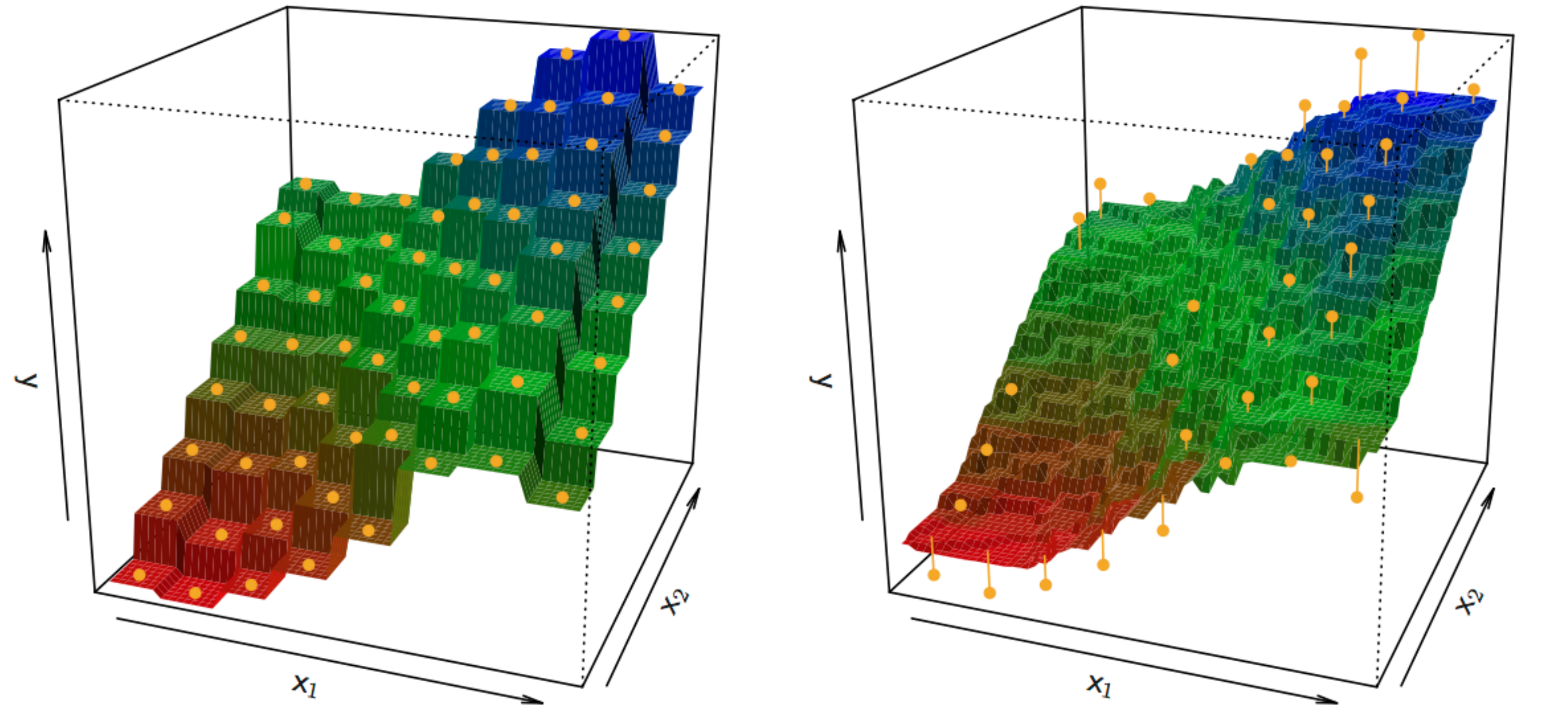
KNN Regression

- kNN Regression is similar to the kNN classifier.
- To predict Y for a given value of X, consider k closest points to X in training data and take the average of the responses. i.e.

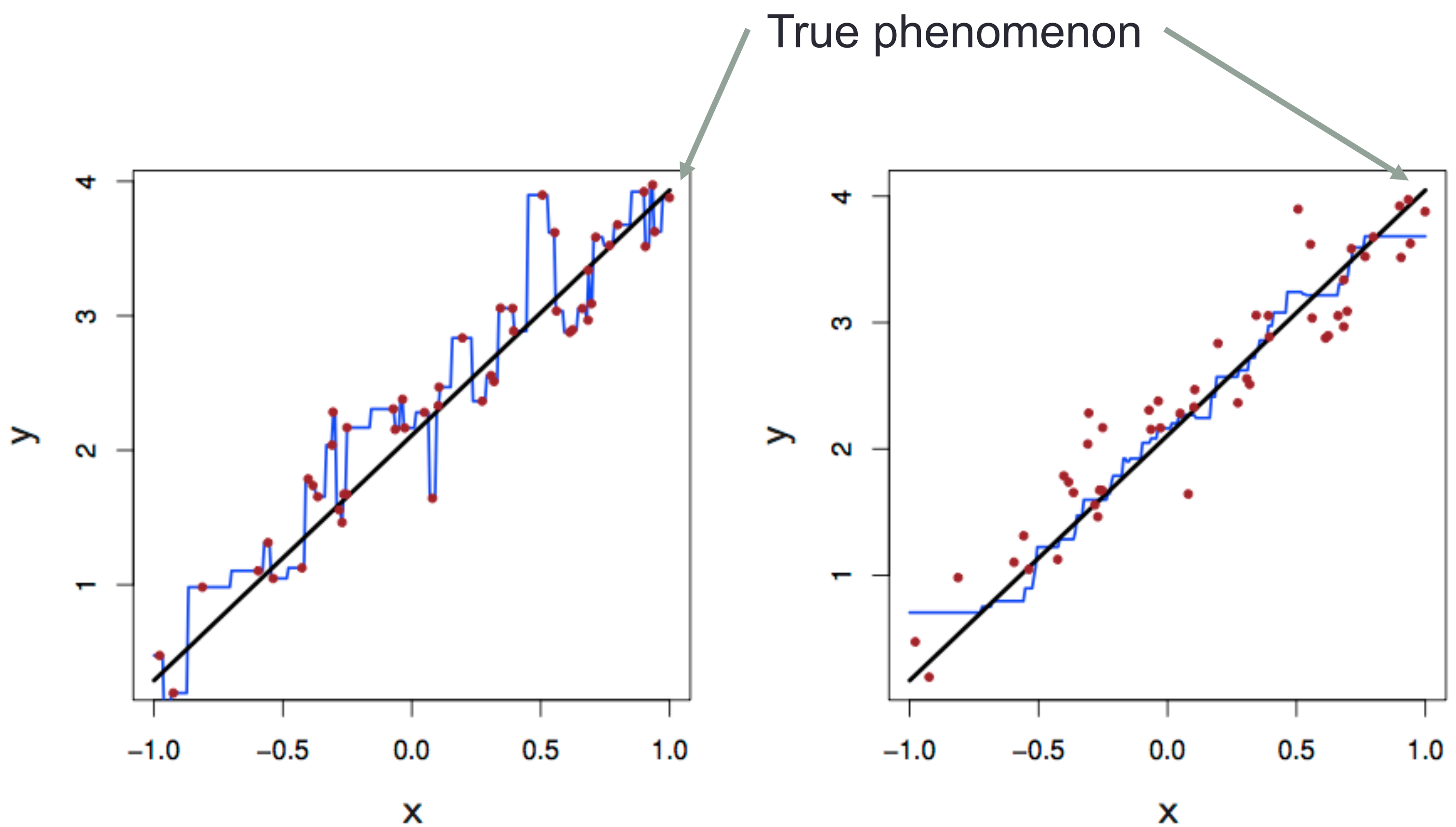
$$f(x) = \frac{1}{K} \sum_{x_i \in N_i} y_i$$

- If k is small kNN is much more flexible than linear regression.
- Is that better?

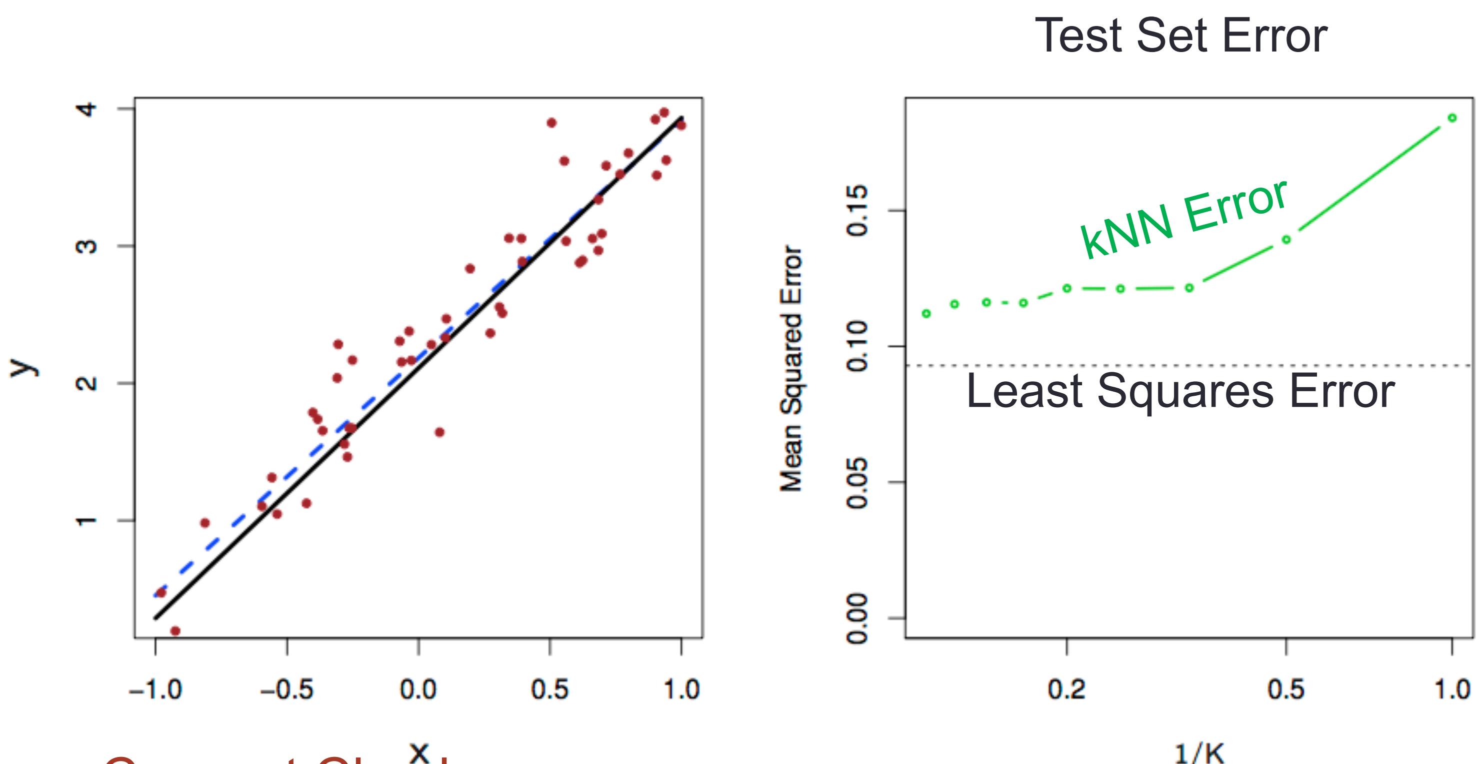
KNN Fits for k =1 and k = 9



KNN Fits in One Dimension (k =1 and k = 9)

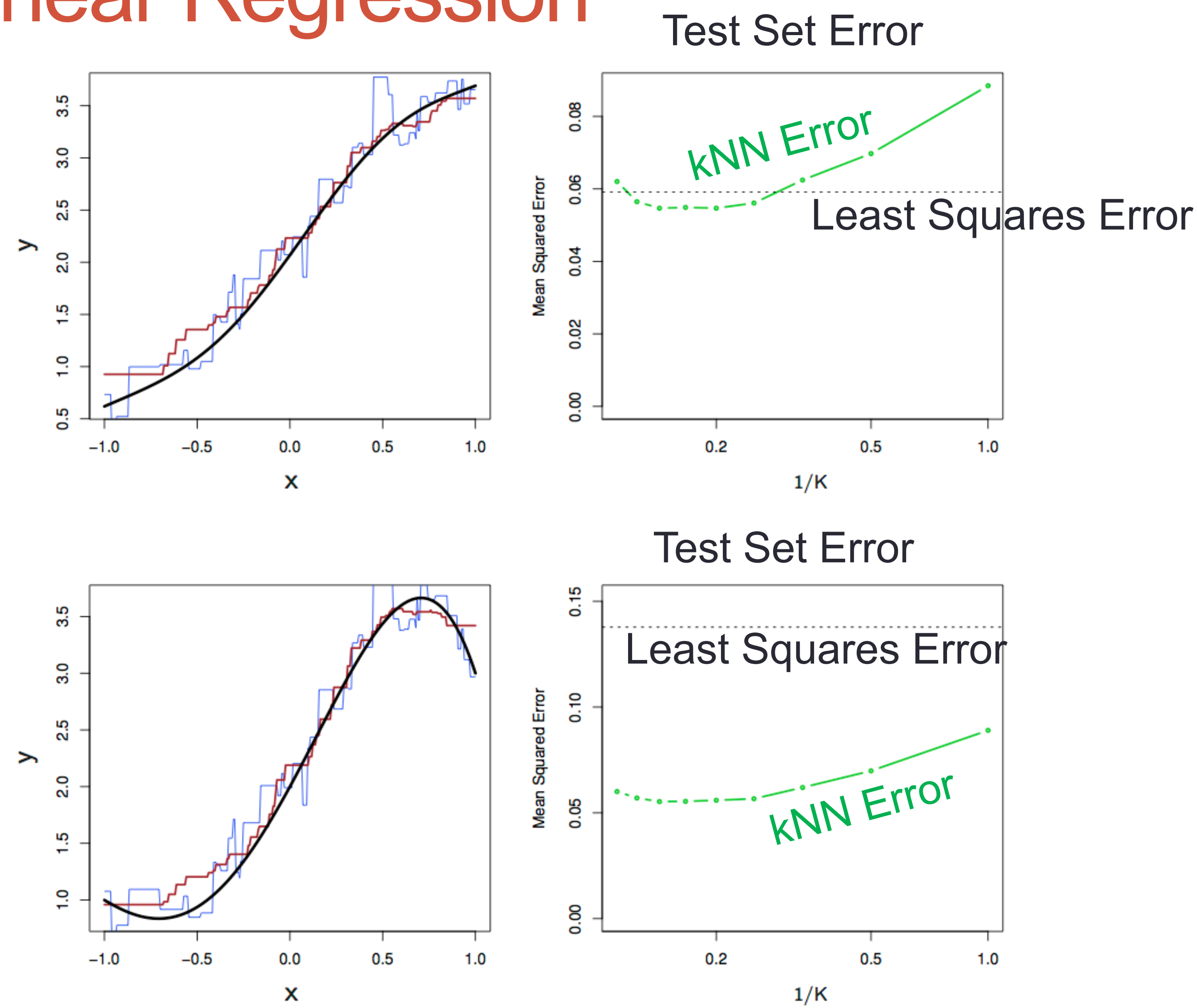


Linear Phenomenon:
Linear Regression Fit vs. kNN

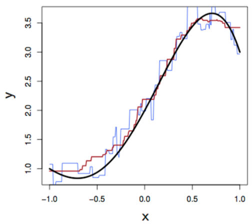


Concept Check:
Why is kNN getting worse as k goes from big to small?

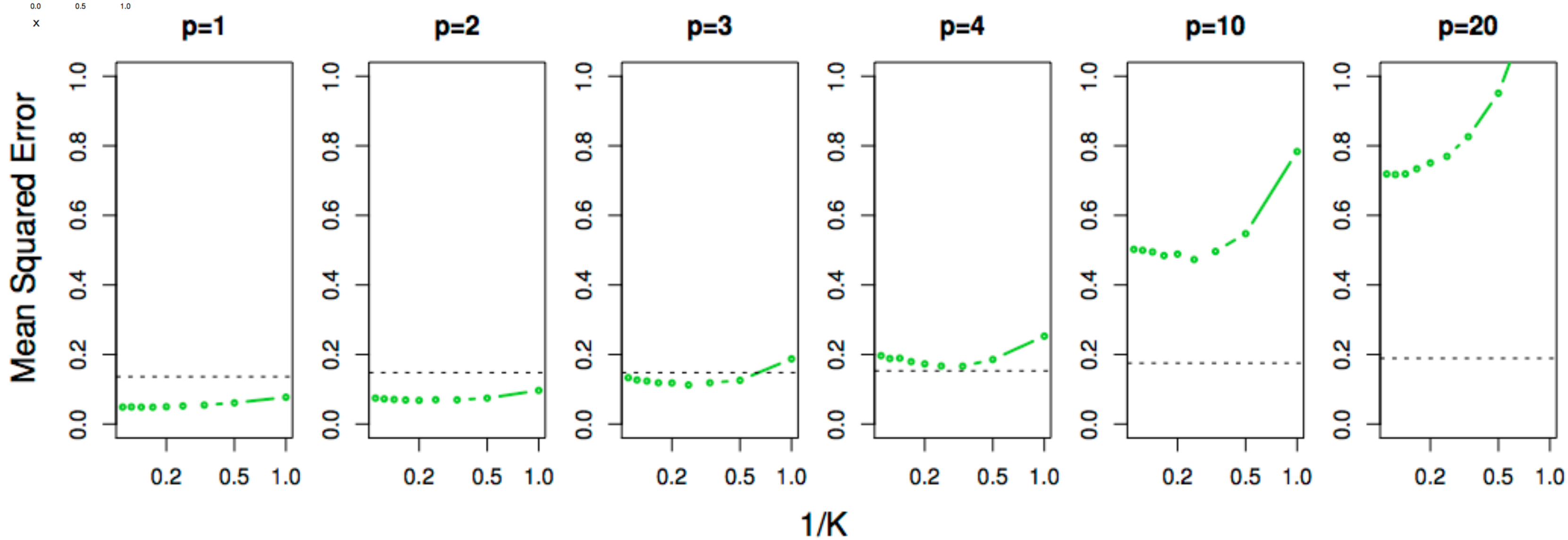
Nonlinear phenomenon: kNN vs. Linear Regression



kNN is Not So Good in High Dimensional Situations



one feature is relevant & nonlinear,
but additional features are irrelevant (noise)



Concept Check: Why does kNN perform ever worse than linear regression as we increase the number of (irrelevant) features?

This behavior is evidence of the phenomenon known as
“The Curse of Dimensionality”