LINEAR REGRESSION

Chapter 03

Outline

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

The Linear Regression Model

$$\boxed{Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon}$$

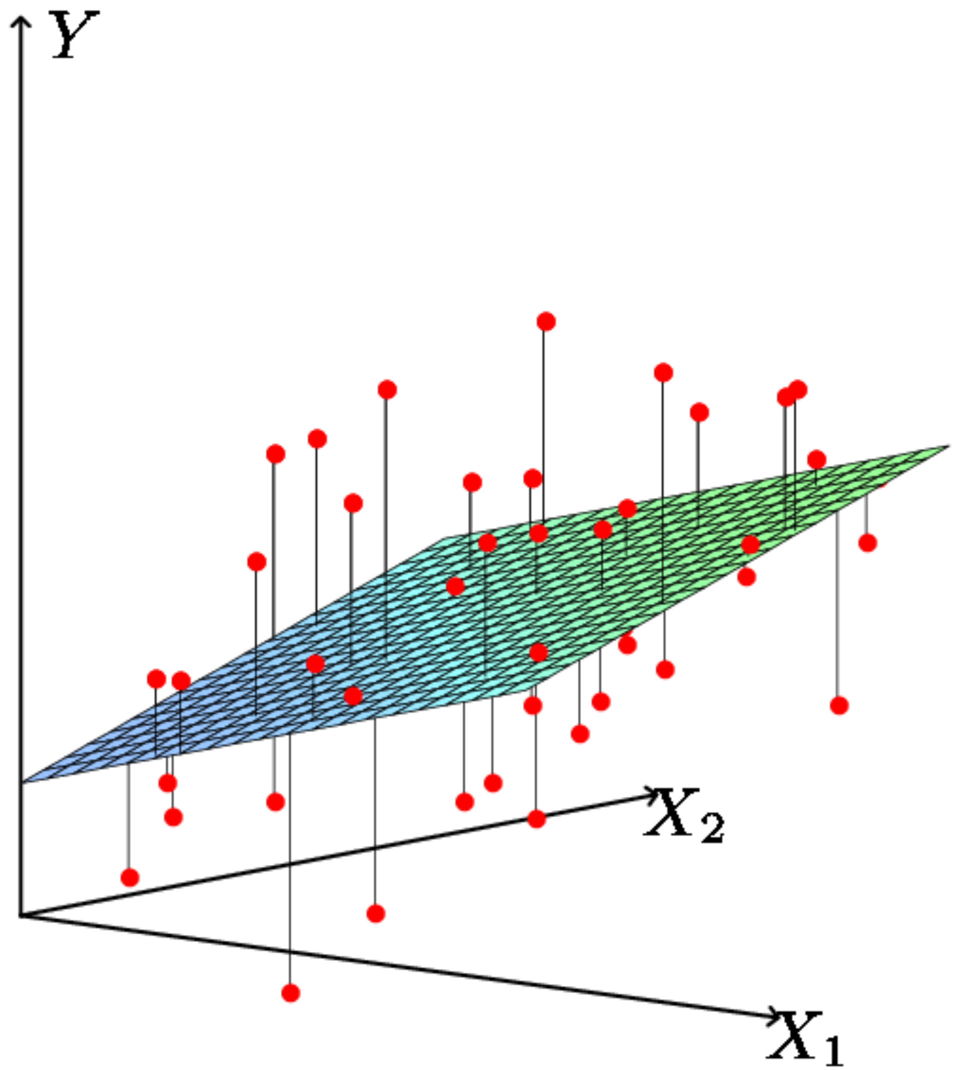
- The parameters in the linear regression model are very easy to interpret.

- 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 jth 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

- We estimate the parameters using least squares i.e. minimize



$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_1 - \cdots - \hat{\beta}_p X_p\right)^2$$

Relationship between population and least squares lines

Population line

$$\boxed{Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon}$$

Least Squares line

$$\boxed{\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p}$$

- We would like to know β_0 through β_p i.e. the population line. Instead we know $\hat{\beta}_0$ through $\hat{\beta}_p$ i.e. the least squares line.
- Hence we 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_i. The guesses will not be perfect just as $\overline{X}$ is not a perfect guess for μ.

Measures of Fit: R^2

- Some of the variation in Y can be explained by variation in the X's and some cannot.

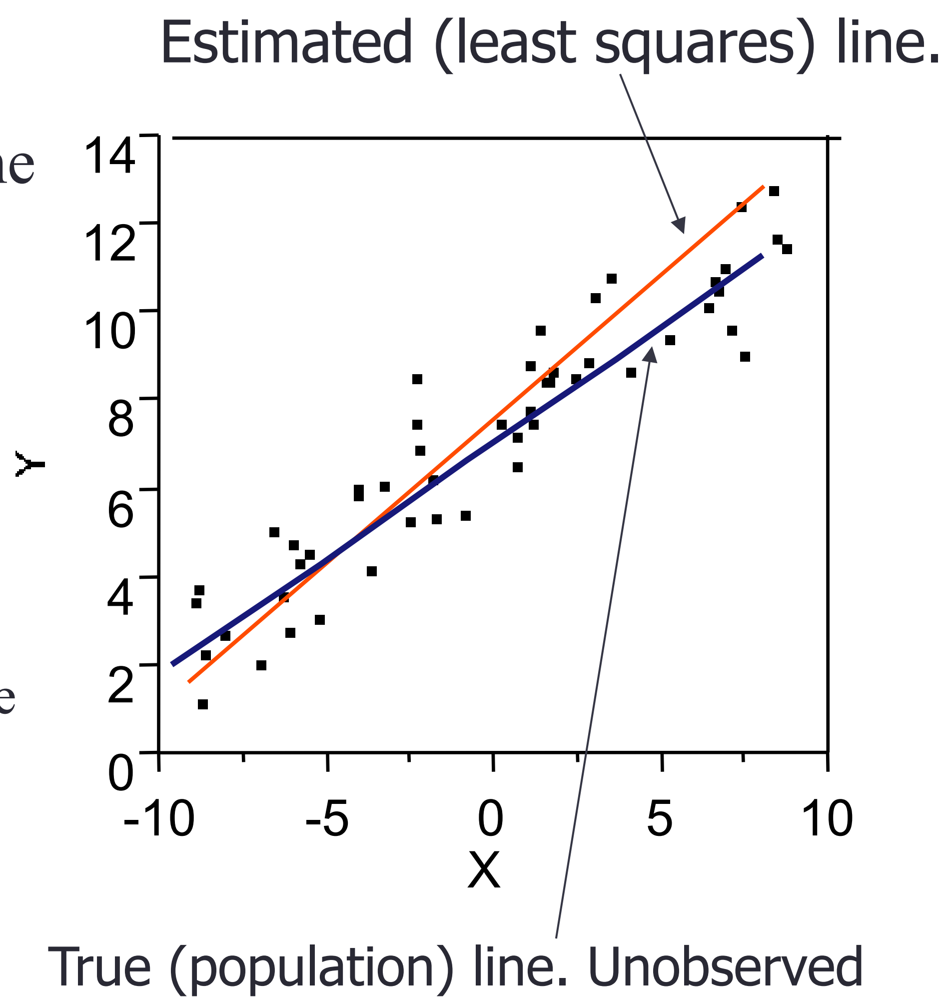
- R^2 tells you the fraction of variance that can be explained by X.

$$R^2 = 1 - \frac{RSS}{\sum (Y_i - \bar{Y})^2} \approx 1 - \frac{\text{Ending Variance}}{\text{Starting Variance}}$$

R^2 is always between 0 and 1. Zero means no variance has been explained. One means it has all been explained (perfect fit to the data).

Inference in Regression

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



Some Relevant Questions

1. Is $\beta_j=0$ or not? We can use a hypothesis test to answer this question. If we can't be sure that $\beta_j \neq 0$ then there is no point in using X_j as one of our predictors.

2. Can we be sure that at least one of our X variables is a useful predictor i.e. is it the case that $\beta_1= \beta_2=\cdots= \beta_p=0$?

1. Is $\beta_j=0$ i.e. 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 $\boxed{t = \hat{\beta}_j / SE(\hat{\beta}_j)}$ ← 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

Is there a (statistically detectable) linear relationship between Newspapers and Sales after all the other variables have been accounted for?

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

No: big p-value

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

Almost all the explaining that Newspapers could do in simple regression has already been done by TV and Radio in multiple regression!

2. Is the whole regression explaining anything at all?

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

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		

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

Outline

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

Qualitative Predictors

- How do you stick “men” and “women” (category listings) into a regression equation?

- Code them as indicator variables (dummy variables)

- For example we can “code” Males=0 and Females= 1.

Interpretation

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

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

- then the regression equation is

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

- β_2 is the average extra balance each month that females have for given income level. Males are the "baseline".

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

Other Coding Schemes

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

$$\boxed{Gender_i = \begin{cases} -1 \text{ if male} \\ 1 \text{ if female} \end{cases}}$$

- then the regression equation is

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

- β_2 is 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.

Other Issues Discussed

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

Interaction

- When the effect on Y of increasing X_1 depends on another X_2.

- Example:
  - Maybe the effect on Salary (Y) when increasing Position (X_1) depends on gender (X_2)?
  - For example maybe Male salaries go up faster (or slower) than Females as they get promoted.

- Advertising example:
  - TV and radio advertising both increase sales.
  - Perhaps 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

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

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

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

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

Parameter Estimates

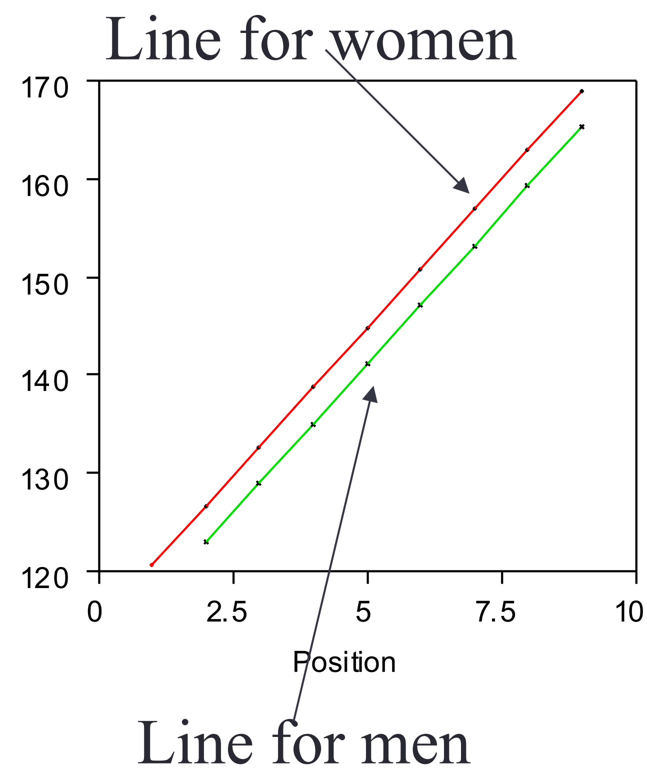
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	6.7502202	0.247871	27.23	<.0001*
TV	0.0191011	0.001504	12.70	<.0001*
Radio	0.0288603	0.008905	3.24	0.0014*
TV*Radio	0.0010865	5.242e-5	20.73	<.0001*

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



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.

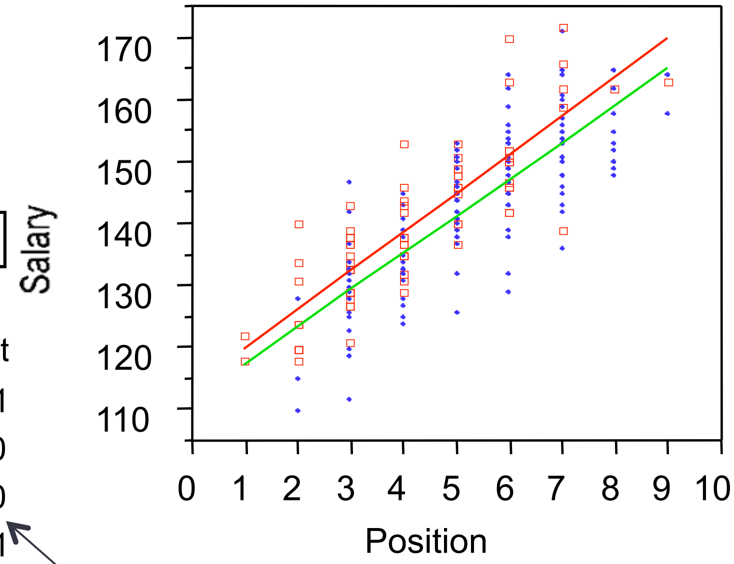
Interaction Effects

- Our model has forced the line for men and the line for women to be parallel.

- Parallel lines say that promotions have the same salary benefit for men as for women.

- If lines aren't parallel then promotions affect men's and women's salaries differently.

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 is not significant

Interaction between gender and position

Outline

- The Linear Regression Model
  - Least Squares Fit
  - 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

There are a number of possible problems that one may encounter when fitting the linear regression model.

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

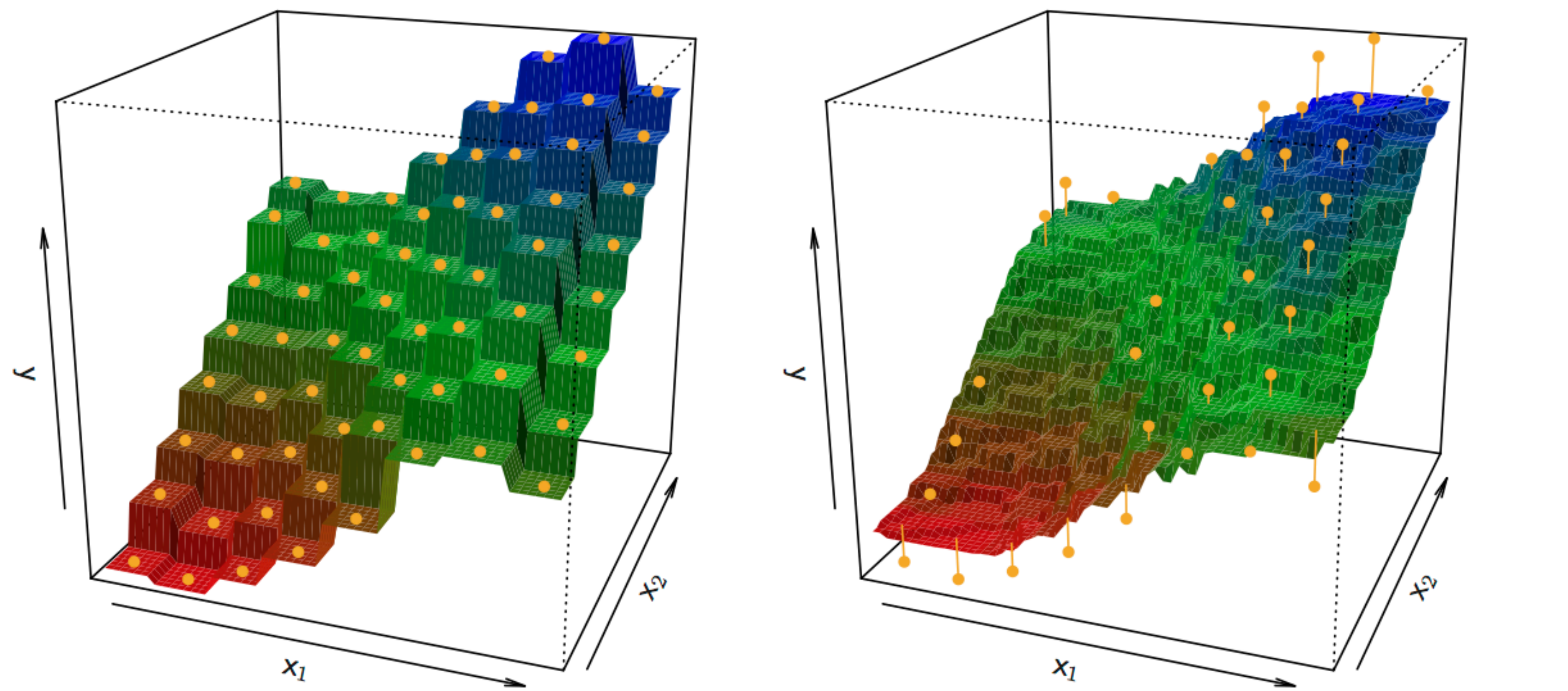
KNN Regression

- 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.

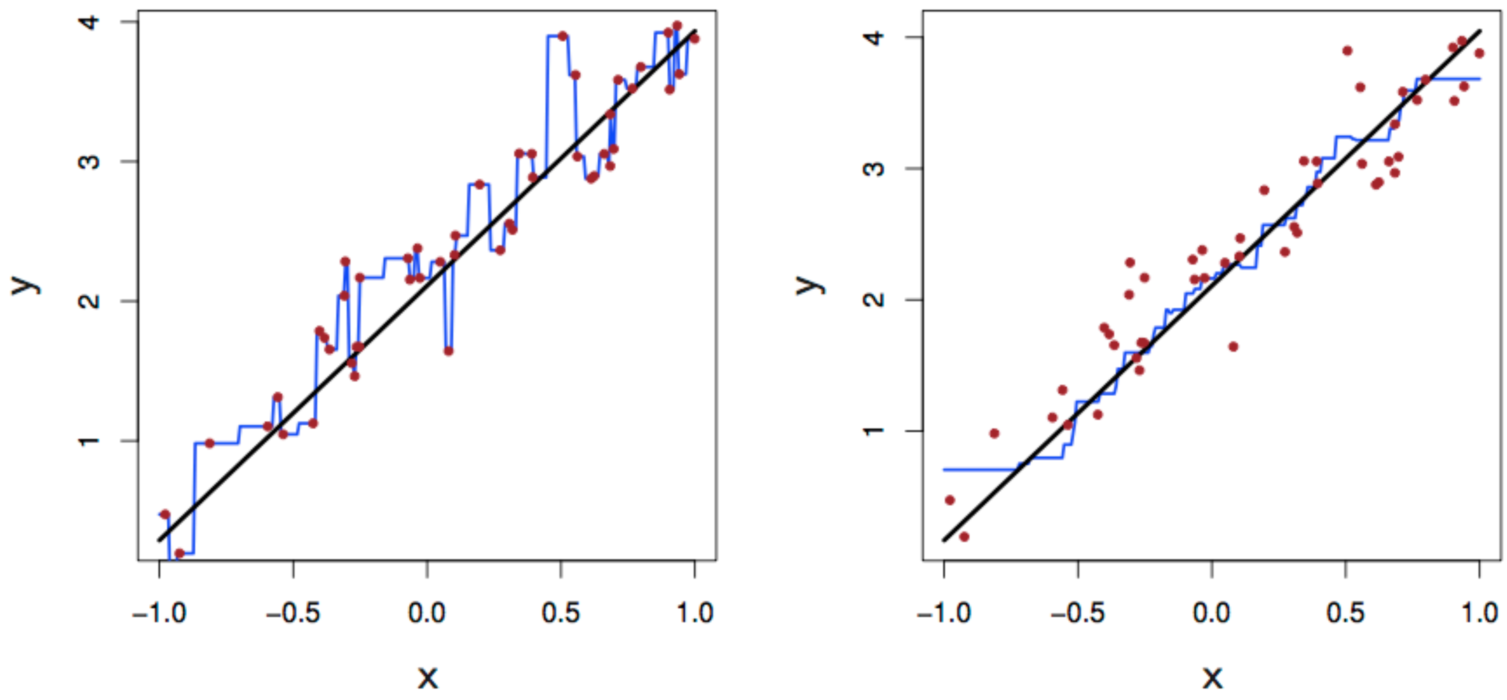
$$f(x) = \frac{1}{K} \sum_{x_i \in N_i} y_i$$

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

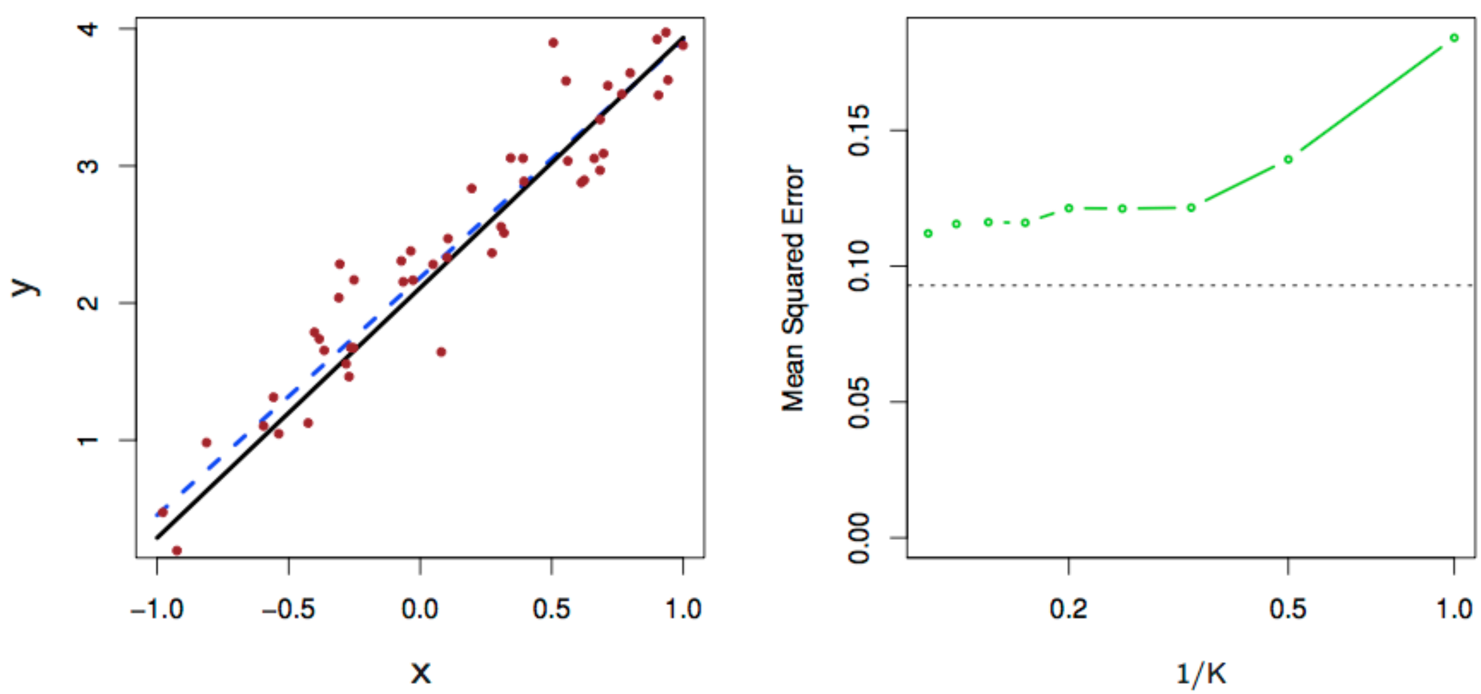
KNN Fits for k =1 and k = 9



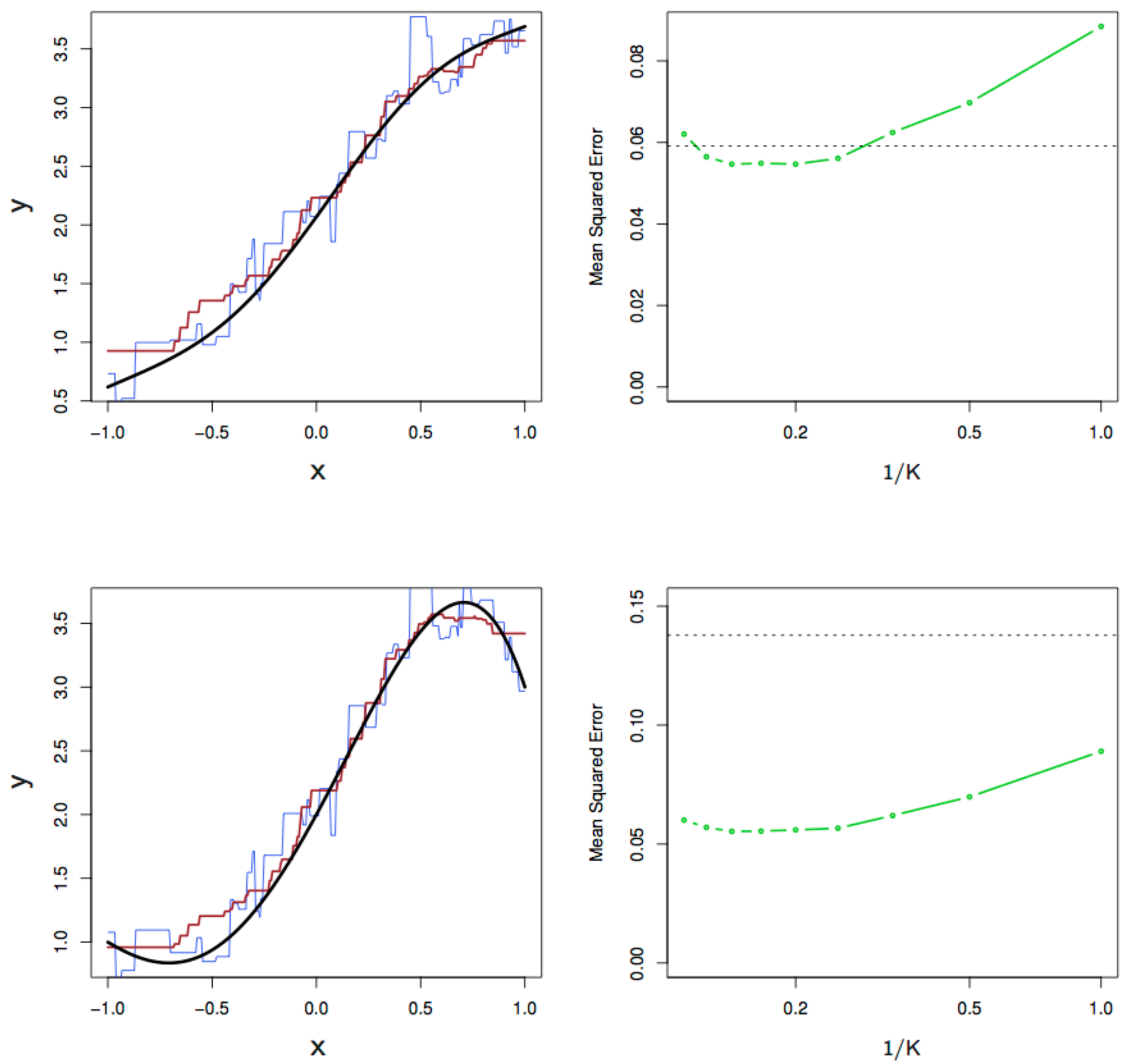
KNN Fits in One Dimension (k =1 and k = 9)



Linear Regression Fit



KNN vs. Linear Regression



Not So Good in High Dimensional Situations