Linear Regression

- Regression models are used in supervised learning when the outcome Y is continuous (numerical) and have in general the form

$$Y = f(X) + \varepsilon$$

Here f is an unknown function, $X^T = (X_1, X_2, \ldots, X_p)$ is the vector of inputs and ε is a random error (independent of X) with mean zero.

- In Linear Regression

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j.$$

Often it's convenient to include the constant variable 1 in X and this way absorb the intercept (or bias) into the vector of parameters β.

- Linear models like the linear regression were largely developed in the pre-computer age of statistics, but there are still good reasons to study and use them
- They are simple and often provide an adequate and interpretable description of how the inputs affect the output.
- In prediction, they can sometimes outperform fancier nonlinear models, especially in situations with small numbers of training cases, low signal-to-noise ratio or sparse data.
- Finally, linear methods can be applied to transformations of the inputs and this considerably expands their scope.
- Here is an outline of the lecture

- The Linear Regression Model
  - Least Squares Fit
  - Measures of Fit
  - Inference in Regression
- Other Considerations in Regression Model
  - Qualitative Predictors
  - Extensions of the Linear Mode – Interaction and Polynomial Terms
- Potential Problems
  - Non-Constant Variance
  - Collinearity etc.
- Comparison to a KNN regression (non-parametric)

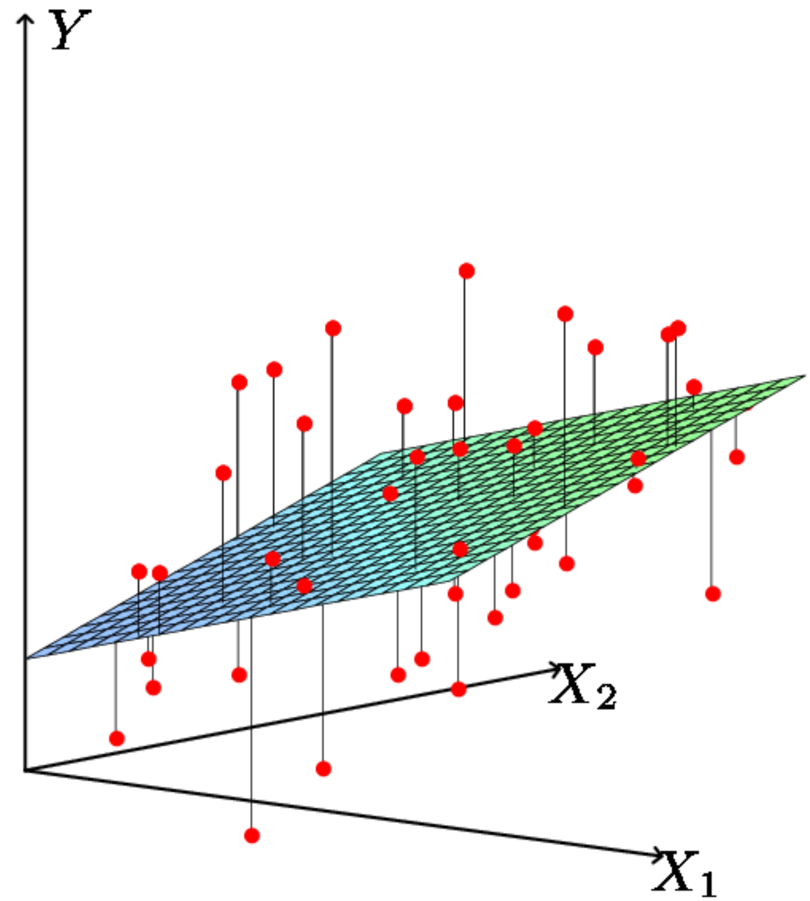
The Linear Regression Model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

- 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

Parameters are estimated by Least Squares, i.e. chosen to minimize the sum of squared residuals

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$
$$= \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip})^2$$



- Formulate the problem in matrix notations as

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{\varepsilon}$$

where $\boldsymbol{y} = (y_1, \dots, y_n)^T$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^T$, $\beta = (\beta_1, \dots, \beta_p)^T$ and

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

- Then $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ are those values of the parameters minimizing

$$\sum_{i=1}^{n} \varepsilon_i^2 = \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} = (\boldsymbol{y} - \boldsymbol{X}\beta)^{\boldsymbol{T}}(\boldsymbol{y} - \boldsymbol{X}\beta)$$

- Differentiating the RHS above w.r.t. β and setting the result to 0 produces the normal equations

$$\boldsymbol{X}^T\boldsymbol{X}\beta = \boldsymbol{X}^T\boldsymbol{y}$$

- Assuming that $\boldsymbol{X}^T\boldsymbol{X}$ is invertible (non-singular) the LS estimator is

$$\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

- This form of the solutions is very convenient to answer questions such as “which predictors are important?” But before that let’s look at the results of a linear regression model fit to one of our model datasets

- For the Advertising data set the result is

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	2.939	0.3119	9.42	< 0.0001
`TV`	0.046	0.0014	32.81	< 0.0001
`radio`	0.189	0.0086	21.89	< 0.0001
`newspaper`	-0.001	0.0059	-0.18	0.8599

TABLE 3.4. *For the* `Advertising` *data, least squares coefficient estimates of the multiple linear regression of number of units sold on radio, TV, and newspaper advertising budgets.*

- The p-values in the last column indicate that `TV` and `radio` are "important" to the outcome `sales`. But `newspaper` is not.

(code)

- But how well does the above LS equation

$$sales \approx 2.939 + 0.046 \times TV + 0.189 \times radio - 0.001 \times newspaper$$

estimate the population law

$$sales = \beta_0 + \beta_1 \times TV + \beta_2 \times radio + \beta_3 \times newspaper + \varepsilon$$

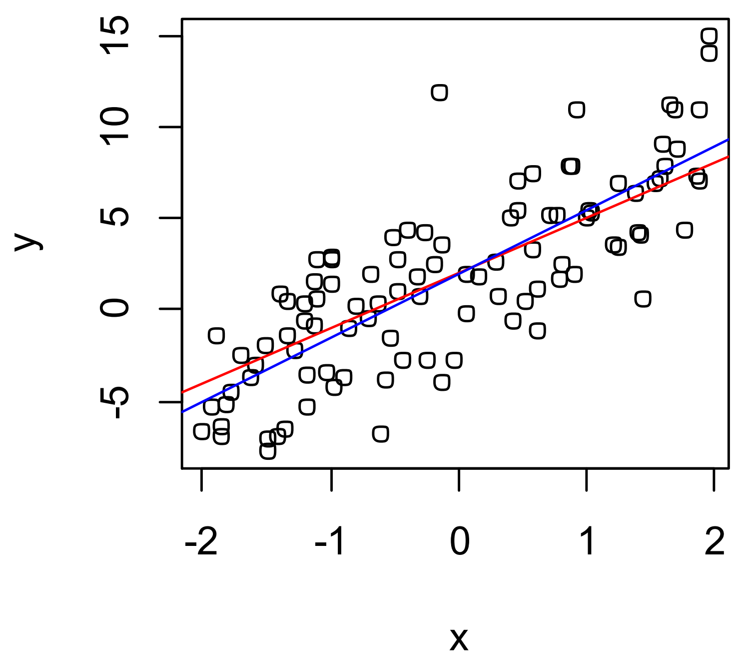
- It's easier to illustrate this question for simple regression ($p = 1$).

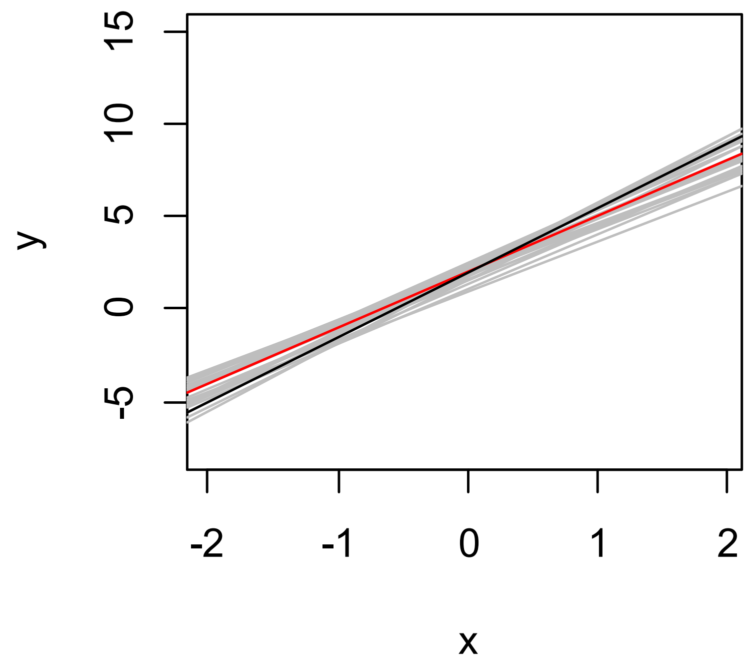
- Create a 100 random Xs and generate 100 corresponding Ys from

$$Y = 2 + 3X + \varepsilon$$

where ε are normally distributed with mean 0.

Plot the true model in red, the LS in blue on left. Then generate randomly 30 datasets from the original data and build LS models. This way we create a distribution for the regression coefficient. Plotting their LS lines in grey we notice they are not far from the true model:





- This observation raises the question of assessing the accuracy of the coefficient estimates and the model fit

- From the general form of the solution, the predicted (fitted) values of Y are

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\beta} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y},$$

$$\boldsymbol{H} \stackrel{\text{def}}{=} \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T \text{ is called } hat\ matrix$$

- Also, the residuals and the residual sum of squares (RSS) are

$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \boldsymbol{X}\widehat{\beta} = \boldsymbol{y} - \hat{\boldsymbol{y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$$

$$\hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{H})^{\boldsymbol{T}}(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y} = \boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y} \quad \text{(RSS)}$$

- Assuming

$$E(\boldsymbol{\varepsilon}) = 0 \ \text{ and } \ \mathrm{var}(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{I}$$

the following important properties of LS regression can be shown

(a) $\hat{\beta}$ is unbiased:

$$E\left(\hat{\beta}\right) = \beta$$

(b) The variance (variance-covariance) matrix of $\hat{\beta}$ is

$$\mathrm{var}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

(c) The (unbiased) estimate for σ^2 is

$$\hat{\sigma}^2 = \frac{\hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}}}{n-p} = \frac{RSS}{n-p}$$

(d) $n-p$ is called the degrees of freedom (dof) of the model. Also:

$$se\left(\hat{\beta}_{i-1}\right) = \hat{\sigma}\sqrt{(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{ii}}$$

- By (a) and (d) the 95% CI for β_{i-1} is approximately (why?)

$$[\hat{\beta}_{i-1} - 1.96\times\mathrm{se}(\hat{\beta}_{i-1}), \hat{\beta}_{i-1} + 1.96\times\mathrm{se}(\hat{\beta}_{i-1})]$$

- For the advertising data, the 95% confidence interval for β_0 is [2.328, 3.55] and the 95% confidence interval for β_1 is [0.043, 0.048]

- Therefore, we can conclude that in the absence of any advertising, sales will, on average, fall somewhere between 2,328 and 3,550 units.

- Furthermore, for each $1,000 increase in television advertising, there will be an average increase in sales of between 43 and 48 units.

Goodness of fit

- We can parcel the total sum of squares (corrected for the mean):

SS_T	**=**	**SS_{reg}**	**+**	**SS_{res}**
(TSS)				(RSS)

$$\Longleftrightarrow \quad \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

- It's used in one common choice of a criterion about how well the model fits the data. This is the coefficient of determination or percentage of variance explained or R^2:

$$R^2 \overset{def}{=} 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\text{RSS}}{\text{TSS}} = \frac{\text{SS}_{reg}}{\text{TSS}}$$

- Its range is $0 \leq R^2 \leq 1$ and values closer to 1 indicated a better fit.

- For simple linear regression $R^2 = r^2$ where r is the (Pearson) *correlation* between *x* (the predictor) and *y* (the response)

When we perform multiple linear regression, we usually are interested in answering a few important questions

1) Is at least one of the predictors $X_1, X_2, \ldots, X_p$ useful in predicting the response?

2) Do all the predictors help to explain Y, or is only a subset of the predictors useful?

3) How well does the model fit the data?

4) Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

1) Is the whole regression explaining anything at all?

- Hypothesis testing
  - 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		

- The answer is derived from the F-test in the ANOVA (ANalysis Of VAriance) table
- What is important in the ANOVA table is the F ratio and the corresponding p-value. In the example above, the p-value is very small indicating that the predictors in the regression "help"

2) Is $\beta_i \neq 0$, i.e. is X_i an important variable?

- We use a hypothesis test to answer this question

$$H_0: \beta_i = 0 \quad \text{vs} \quad H_a: \beta_i \neq 0$$

- Calculate

$$\boxed{t_i = \hat{\beta}_i / se(\hat{\beta}_i)}$$

- If t_i is large (equivalently p-value is small) we can be sure that $\beta_i \neq 0$ and that there is a relationship

- In the simple regression of `sales` on `TV`, the `TV` predictor is important

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

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

Yes (small p-value)

- Almost all the explaining that `newspapers` could do in simple regression has already been done by `TV` and `radio` in multiple regression!

Qualitative Predictors

- Consider dataset Credit
- How do you stick “men” and “women” (category listings) into a regression equation predicting balance?
- 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

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

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

More Coding Schemas

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

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

- Then the regression equation is

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

- Now, 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 Important Considerations

- Interaction terms
- Non-linear effects
- Collinearity
- Model Selection – will consider it in a few lectures

Interaction

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

- Example:
  - Maybe the effect on Salary (Y) when changing Position (X_1) depends on Gender (X_2)?
  - For example, `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 data

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

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*

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

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

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

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

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

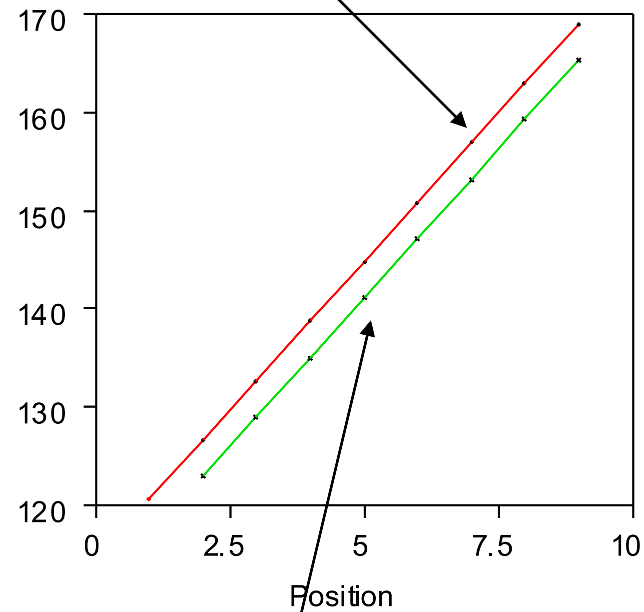
female: salary = $\boxed{112.77+1.86}$ + **6.05**× position

males: salary = $\boxed{112.77-1.86}$ + **6.05** × position

Different intercepts

Same slopes

Line for women



Line for men

- 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.

Most Common Potential Problems

1. Non-linearity of the response-predictor relationships.
2. Correlation of error terms.
3. Non-constant variance of error terms.
4. Outliers.
5. High-leverage points.
6. Collinearity

Examples [ISLR]:

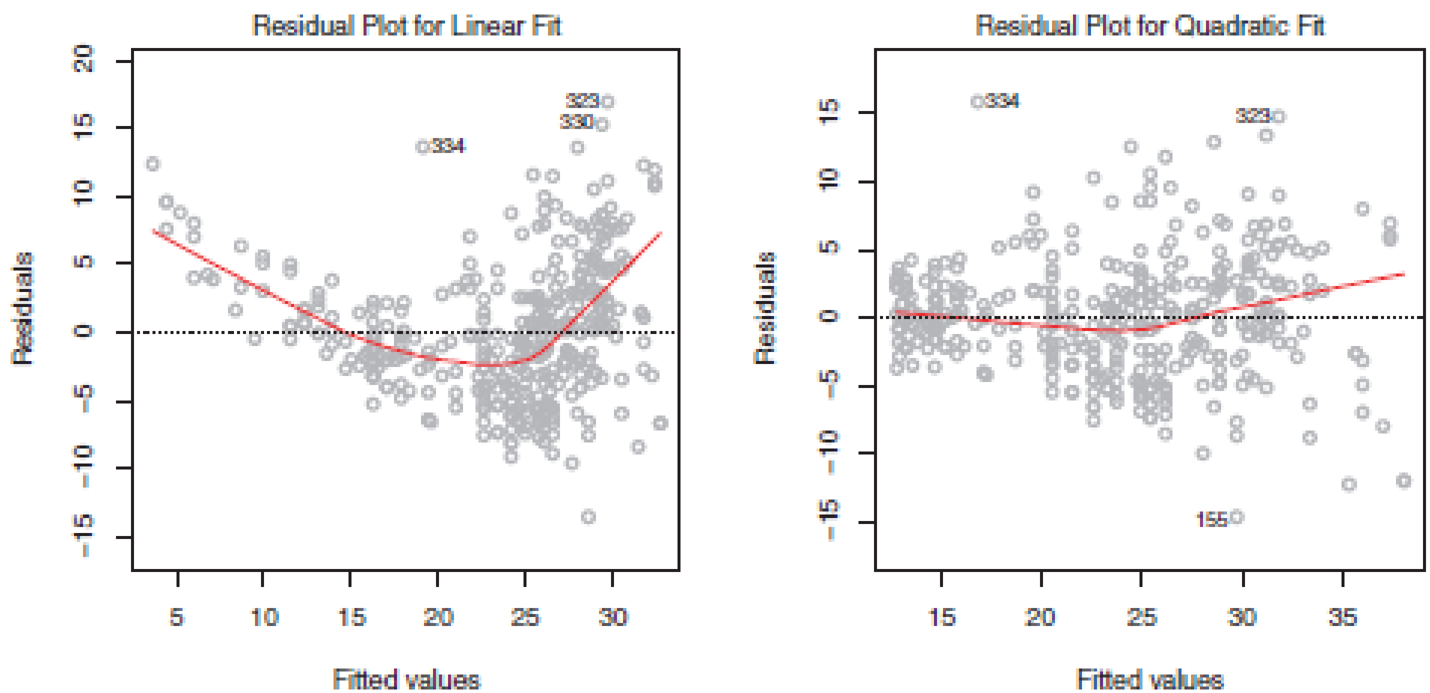


FIGURE 3.9. *Plots of residuals versus predicted (or fitted) values for the* Auto *data set. In each plot, the red line is a smooth fit to the residuals, intended to make it easier to identify a trend.* Left: *A linear regression of* mpg *on* horsepower*. A strong pattern in the residuals indicates non-linearity in the data.* Right: *A linear regression of* mpg *on* horsepower *and* horsepower^2*. There is little pattern in the residuals.*

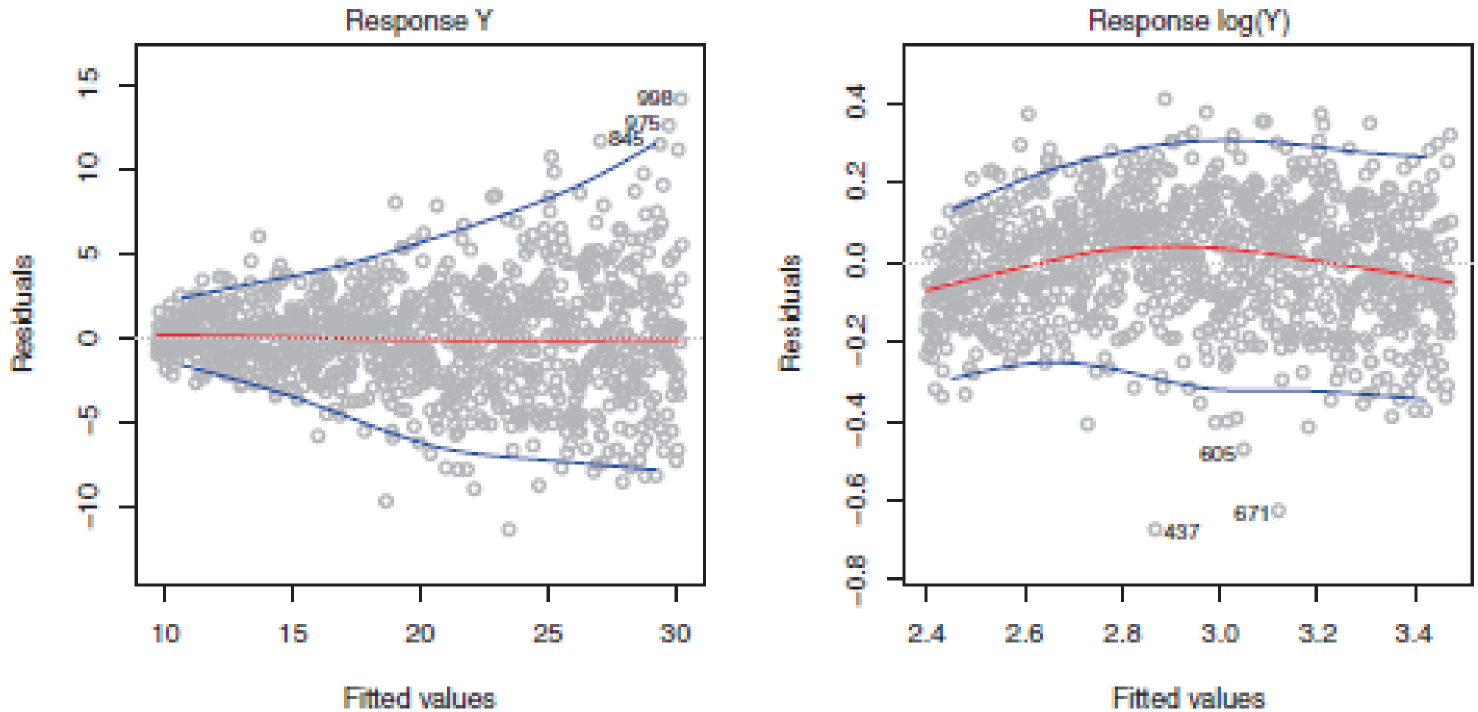


FIGURE 3.11. *Residual plots. In each plot, the red line is a smooth fit to the residuals, intended to make it easier to identify a trend. The blue lines track the outer quantiles of the residuals, and emphasize patterns.* Left: *The funnel shape indicates heteroscedasticity.* Right: *The response has been log transformed, and there is now no evidence of heteroscedasticity.*

General linear regression model building strategy

1) Plot and summarize your data and inspect it.

2) Fit a model to your data. Check the fit

3) Run **diagnostics** to check assumptions:

a) constant variance; b) linearity; c) normality; d) outliers; e) influential points; f) serial correlation and g) colinearity

4) **Transform**:

a) Box-Cox for the response

b) polynomial regressions, $log(\,)$ for the predictors

5) **Variable selection**: testing- and criterion-based methods

Repeat the steps if necessary but be aware of *too much* analysis. So, avoid complex models for small datasets.

Comparison to Non-Parametric Methods

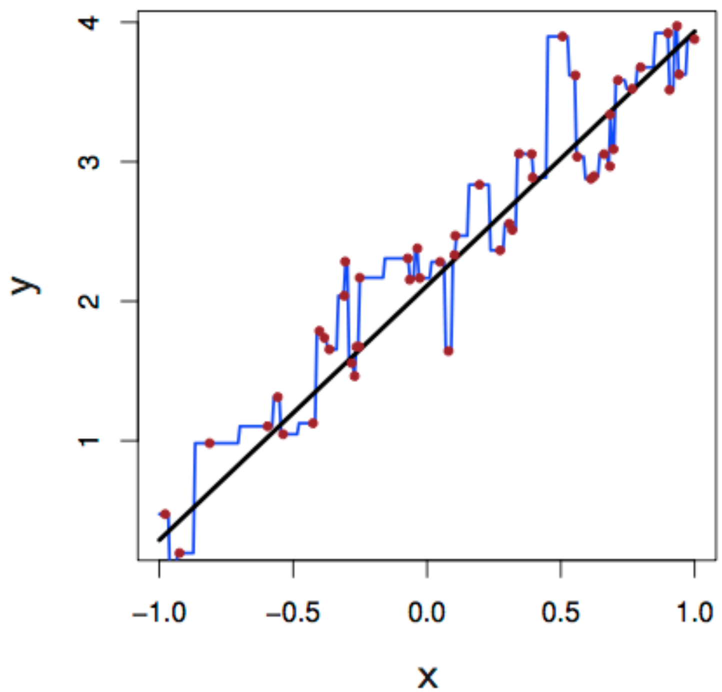
- The nearest neighbor methods use those observations closest in the input space to a point x_0 to form the prediction. Euclidean distance is regularly used to determine proximity

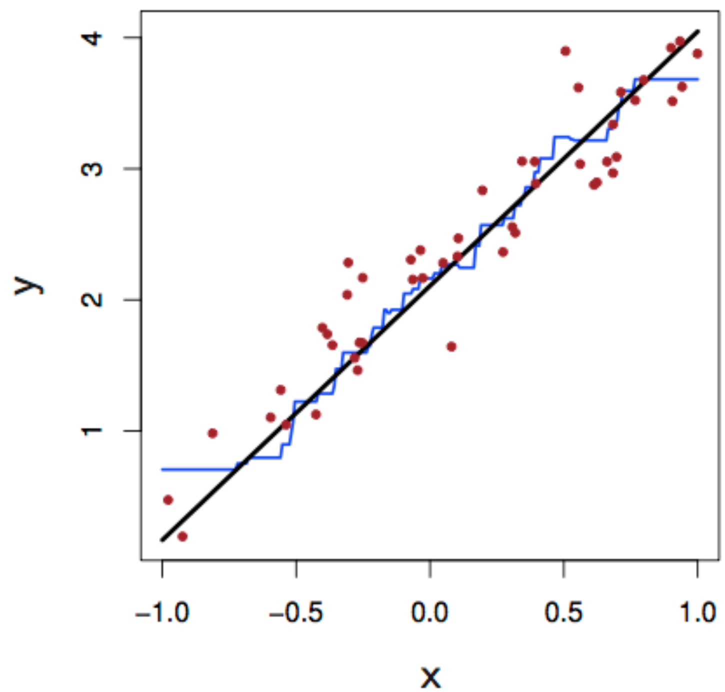
- Given a value for K and a prediction point x_0, KNN regression first identifies the K training observations that are closest to x_0, represented by $N_k(x_0)$

- It then estimates $f(x_0)$ using the average of all the training responses in $N_k(x_0)$

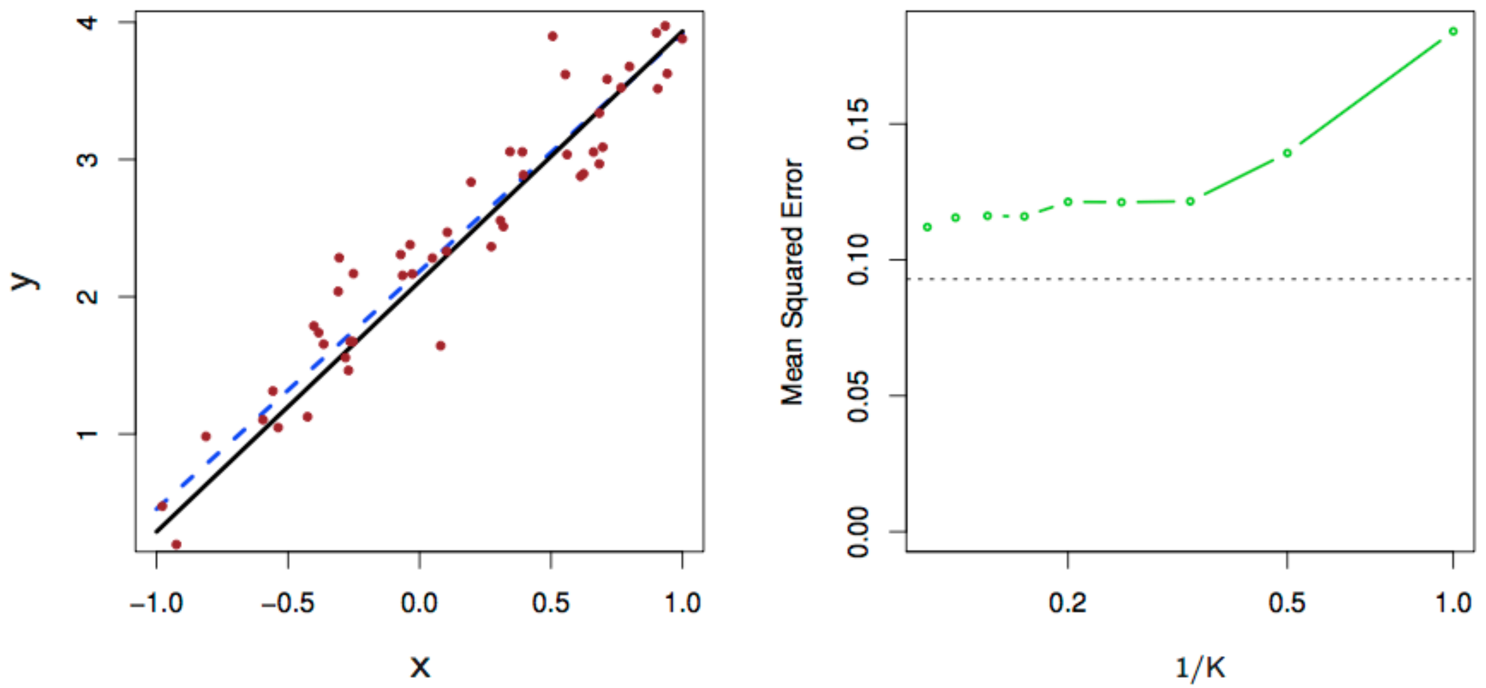
$$\hat{f}(x_0) = \frac{1}{K} \sum_{x_i \in N_0} y_i$$

KNN in 1 dimension ($k = 1$ and $k = 9$)





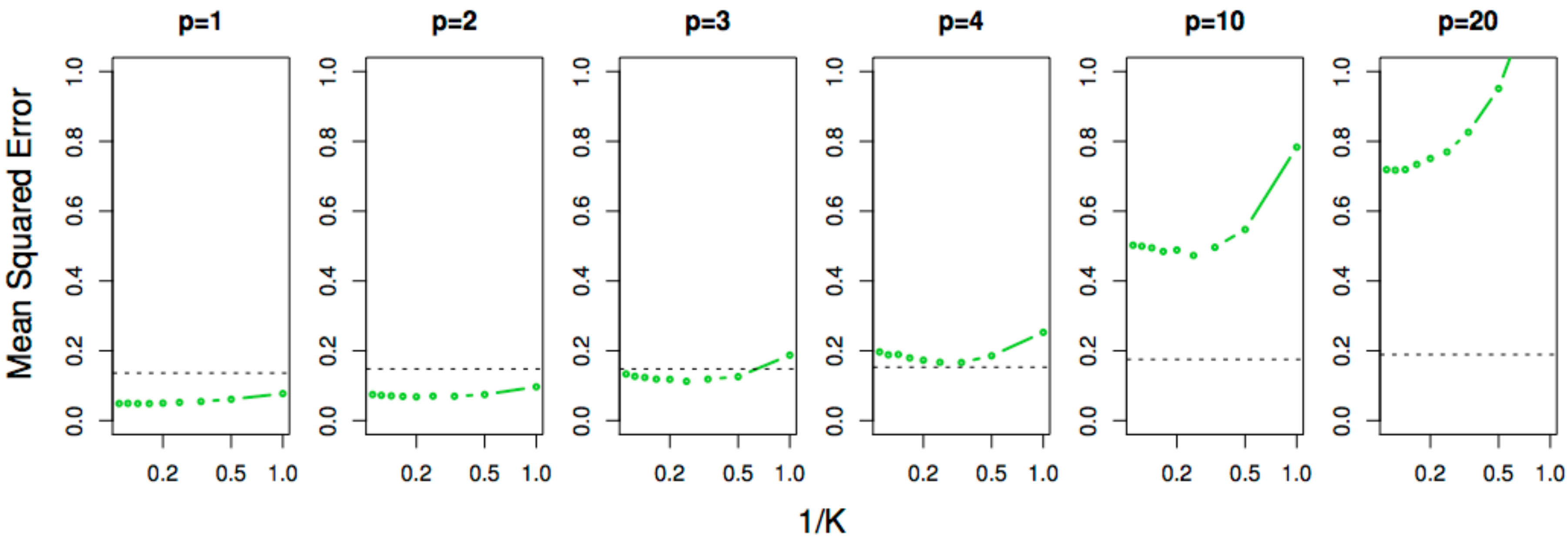
The black line is the linear regression.



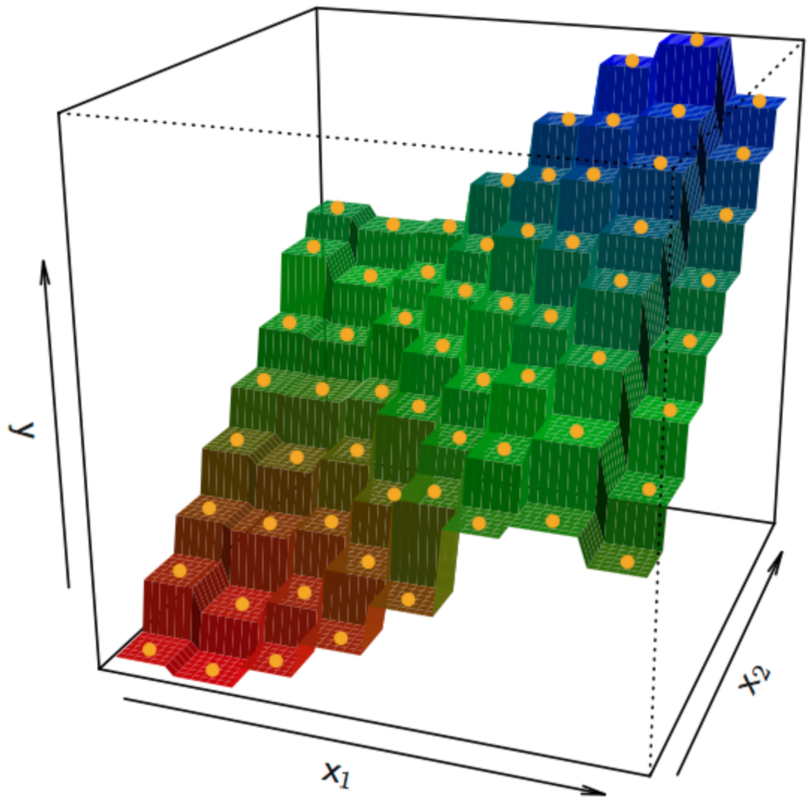
Left: Population line (black), LS fit (dashed blue)

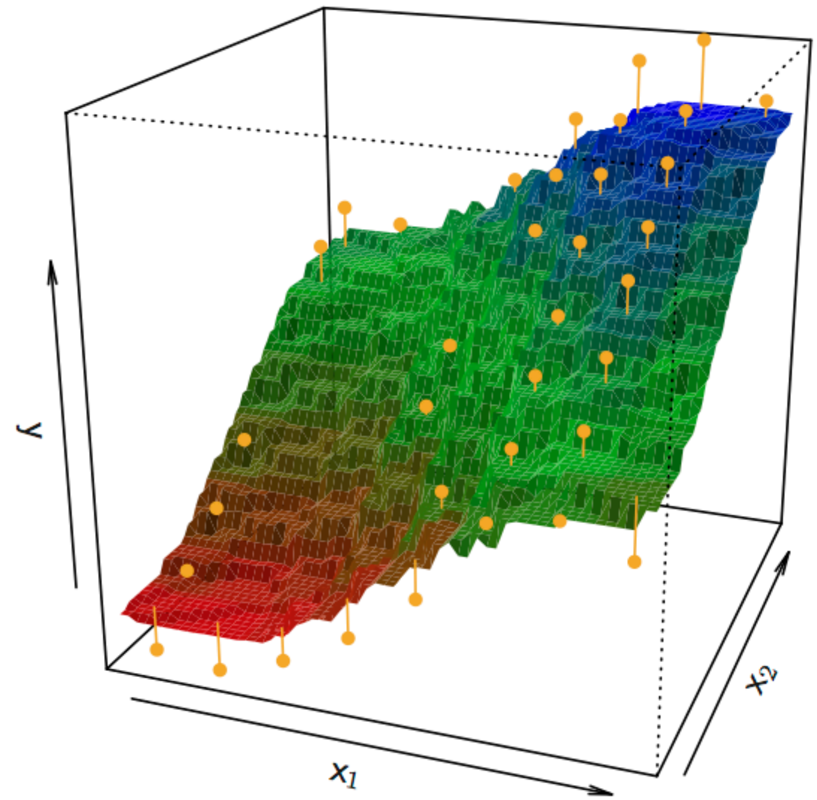
Right: Irreducible error (horizontal line), Green (number of neighbors k)

- One reason KNN is not often used for regression is the "curse of dimensionality"
- Below, the test MSE of a linear regression is displayed as a dashed line and green line is the KNN's test MSE as a function of k
- With the number of predictors p increasing, clearly the linear regression starts to behave better. This is because, if n is not very large, the closest points to x_0 are further away when the predictor space dimension increases



2-dimensions: KNN Fits for $k = 1$ and $k = 9$





Statistical Decision Theory

- $X \in R^p$ – **random** input vector and $Y \in R$ – **random** output variable have joint distribution $\Pr(X, Y)$
- $f(X)$ – a function we are looking for to use to predict Y
- $L(Y, f(X))$ – a loss function used to penalize for the error of prediction
- Common choice is the squared loss function

$$L\big(Y, f(X)\big) = (Y - f(X))^2$$

- We choose f according to the expected (squared) prediction error (EPE)

$$EPE(f) = E(Y - f(X))^2 = \int [y - f(x)]^2 \Pr(dx, dy)$$

- Then apply the iterated expectation formula (conditioning on X)

$$EPE(f) = E_X E_{Y|X}([Y - f(X)]^2|X)$$

which can be minimized w.r.t. f pointwise

$$f(x) = \text{argmin}_c \, E_{Y|X}([Y - c]^2|X = x)$$

- The solution is the conditional expectation also called the regression function

$$f(x) = E(Y|X = x)$$

- The KNN method directly implements the above recipe (on training data). At each point x_0 we want to average all those y_i's with input $x_i = x_0$. Typically there is only one observation for each x, so we are forced to use

$$\hat{f}(x_0) = \text{Ave}\,(y_i|x_i \in N_K(x_0))$$

where “Ave" denotes average.

Two approximations are happening here:

- Expectation is approximated by averaging over sample data

- Conditioning at a point is relaxed to conditioning on some region "close" to the target point.

For large training sample size N, the points in the neighborhood are likely to be close to x_0, and as k gets large the average will get more stable.

In fact, under mild conditions on the joint probability distribution $\Pr(X, Y)$, one can show that as $N, k \to \infty$ such that $k/N \to 0$, $\hat{f}(x) = E(Y|X = x)$

Why not use the KNN in all problems then? There are 2 main reasons for this

- Sample size
- Curse of the dimensionality

So both k-nearest neighbors and least squares end up approximating conditional expectations by averages. But they differ dramatically in terms of model assumptions:

- Least squares assumes $f(x)$ is well approximated by a globally linear function.
- k-nearest neighbors assumes $f(x)$ is well approximated by a locally constant function

Reading:

ESLII: Chapter 2 (but specifically 2.1-2.4)

ISLR: Chapter 3