Linear Regression [1]

Shaobo Li

University of Kansas

[1] Partially based on Hastie, et al. (2009) ESL, and James, et al. (2013) ISLR

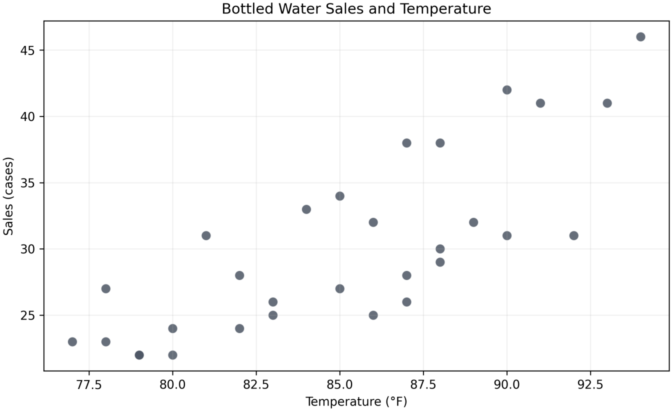
Linear regression – a fundamental learning algorithm

- Supervised learning method
- It assumes the dependence of Y on X is linear
- Largely used in many disciplines
- Simple and interpretable
- Fundamental in data science

What can linear regression do?

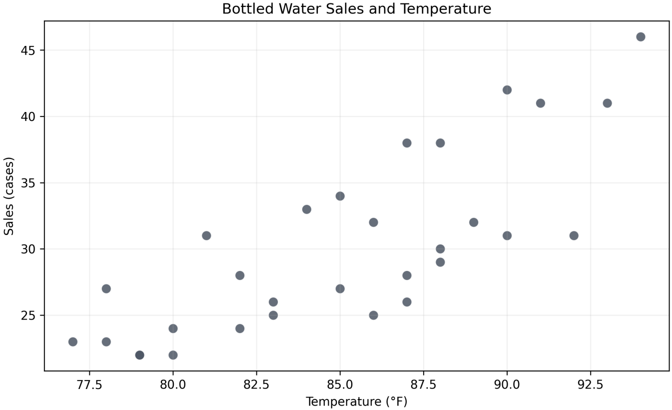
- Is there an association between X and Y?
- If yes, how strong is this association?
- Is this association linear?
- If there are multiple X's (X_1, X_2 and X_3), which of them are related to Y and which are not?
- Can we predict the value of Y for any given X?
- How accurate is such prediction?

An example – bottled water sales



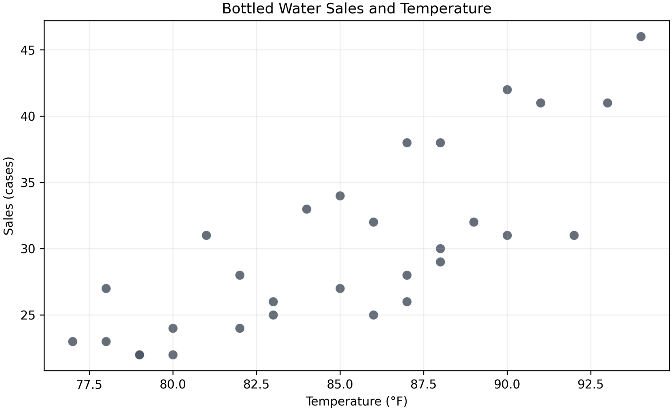
- Is there an association between *Temperature* and *Sales*?

An example – bottled water sales



- Is there an association between *Temperature* and *Sales*?
- If yes, how strong is this association?

An example – bottled water sales



- Is there an association between *Temperature* and *Sales*?
- If yes, how strong is this association?
- Is this association linear?

An example – bottled water sales

- We can write this relationship as

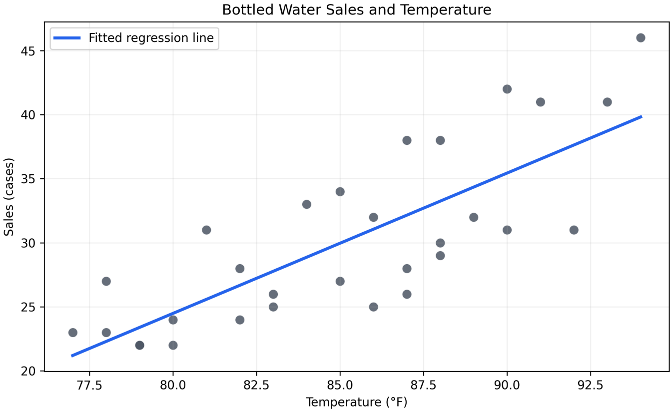
$$Sales \approx \beta_0 + \beta_1 \times Temperature$$

- We use "$\approx$" because the model always approximates the "truth".
- This is a *simple linear regression*.
- β_0 is called intercept and β_1 is slope.
- Given the data points we observed, the model is estimated to be

$$Sales \approx -63.11 + 1.09 \times Temperature$$

- How do we interpret this model?

Fitted straight line – a simple linear regression



Linear regression models

- More generally, a simple linear regression model is

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

and a multiple linear regression model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon,$$

where ϵ catches the error of the "model" from the "truth".

- Y is called dependent variable (or response, outcome).
- X is called independent variable (or covariates, explanatory variable).

Linear regression model in matrix form

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1p} \\ 1 & X_{21} & X_{22} & \dots & X_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$\boldsymbol{y} = \mathbf{X}\beta + \boldsymbol{\epsilon}$$

- $\mathbf{X}$ is called design matrix

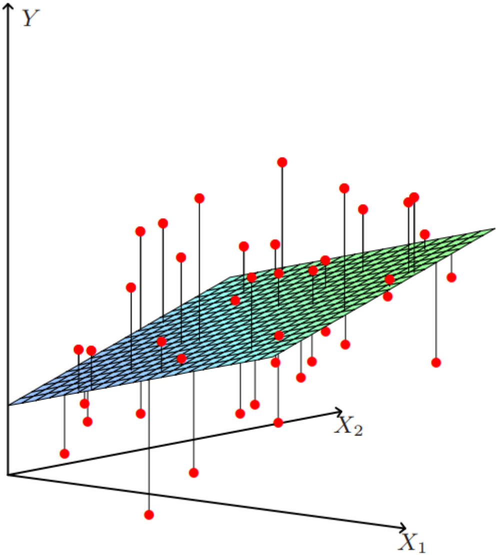
Model estimation

- The *estimated linear regression model* is

$$\hat{\boldsymbol{y}} = \mathbb{E}(\boldsymbol{y}|\mathbf{X}) = \mathbf{X}\hat{\beta}$$

- We need to figure out $\hat{\beta}$, the estimates of β
- Method to use: *ordinary least square (OLS)*

Least square solution



Least square solution

- We want to minimize residual sum squares (RSS)

$$RSS(\beta) = \sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\beta)^2$$
$$= (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)$$

- Take first-order derivative with respect to β and set to 0

$$0 = \frac{\partial RSS(\beta)}{\partial \beta} = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta)$$
$$\mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{X}\beta$$

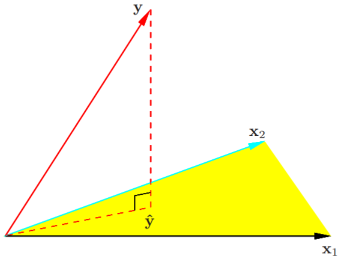
- This is called *normal equation*.

Least square solution

- By assuming $p < n$, the OLS solution is

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

- The predicted value is $\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$
- $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is called hat matrix or projection matrix
- That is, $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$. In other words, $\hat{\mathbf{y}}$ is a linear projection of $\mathbf{y}$



Some important questions after fitting the model

- How to interpret the model?
- Is at least one of the predictors useful to predict and explain the response?
- Do all predictors help to explain response, or just a subset?
- How well does the model fit the data?
- Given a set of predictor values, what response value does the model predict? How accurate is the prediction?

Hypothesis testing – test multiple coefficients

Is at least one X useful?

- F-test for overall significance
  - H_0: $\beta_1 = \ldots = \beta_p = 0$; H_1: at least one $\beta \neq 0$
  - F statistics

$$F^* = \frac{(TSS - RSS)/p}{RSS/(n-p-1)} \sim F_{p,n-p-1}$$

where $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$, is total sum squares

Hypothesis testing – test individual coefficients

Is a specific X relevant?

- Testing for individual β
  - H_0: $\beta_j = 0$; H_1: $\beta_j \neq 0$
  - Using T-test since the true variance is unknown

$$T = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{v_j}} \sim t_{n-p-1}$$

  where v_j is the jth diagonal element of $(\mathbf{X}^T\mathbf{X})^{-1}$
  - Reject H_0 if p-value $< \alpha$ or $|T| > T_{1-\alpha}^{(n-p-1)}$
- Confidence interval: $\hat{\beta} \pm se(\hat{\beta}) \times T_{1-\alpha}^{(n-p-1)}$

R output for bottle water example

```
> model1 <- lm(Sales~Temperature, data = water)
> summary(model1)

Call:
lm(formula = Sales ~ Temperature, data = water)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.6253 -2.9128 -0.9378  4.1109  6.5647 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -63.1146    13.7045  -4.605 8.15e-05 ***
Temperature   1.0950     0.1609   6.807 2.15e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.145 on 28 degrees of freedom
Multiple R-squared:  0.6233,	Adjusted R-squared:  0.6099 
F-statistic: 46.34 on 1 and 28 DF,  p-value: 2.148e-07
```

Python output for bottle water example

```python
import statsmodels.formula.api as smf
temperature_model = smf.ols("Sales ~ Temperature", data=water).fit()
print(temperature_model.summary())
```

✓ 57ms | 9 mins ago

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Sales   R-squared:                       0.623
Model:                            OLS   Adj. R-squared:                  0.610
Method:                 Least Squares   F-statistic:                     46.34
Date:                Wed, 09 Sep 2026   Prob (F-statistic):           2.15e-07
Time:                        13:34:29   Log-Likelihood:                -84.187
No. Observations:                  30   AIC:                             172.4
Df Residuals:                      28   BIC:                             175.2
Df Model:                           1
Covariance Type:            nonrobust
===============================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------
Intercept     -63.1146     13.705     -4.605      0.000     -91.187     -35.042
Temperature     1.0950      0.161      6.807      0.000       0.765       1.425
==============================================================================
Omnibus:                        6.685   Durbin-Watson:                   0.983
Prob(Omnibus):                  0.035   Jarque-Bera (JB):                2.036
Skew:                           0.113   Prob(JB):                        0.361
Kurtosis:                       1.744   Cond. No.                     1.54e+03
==============================================================================
```

Model Assessment – R Square and MSE

- It is proportion of variation in Y explained by the model

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

  R^2 increases monotonically as number of X's increasing.

- Adjusted R^2

$$R^2_{adj} = 1 - \frac{n-1}{n-p-1}\frac{RSS}{TSS}$$

- Mean Squared Error (MSE)

$$MSE = \frac{1}{n-p-1} \times RSS$$

  It is an unbiased estimate of σ^2, variance of ϵ. (Can you show this?)

Categorical covariate

Now consider another variable "Game Day".

$$\text{Game Day} = \begin{cases} 1 & \text{game day} \\ 0 & \text{non-game day} \end{cases}$$

- Sales (Y); Temperature (X_1); Game Day (X_2)
- How does the linear model look like?

Categorical covariate

Now consider another variable "Game Day".

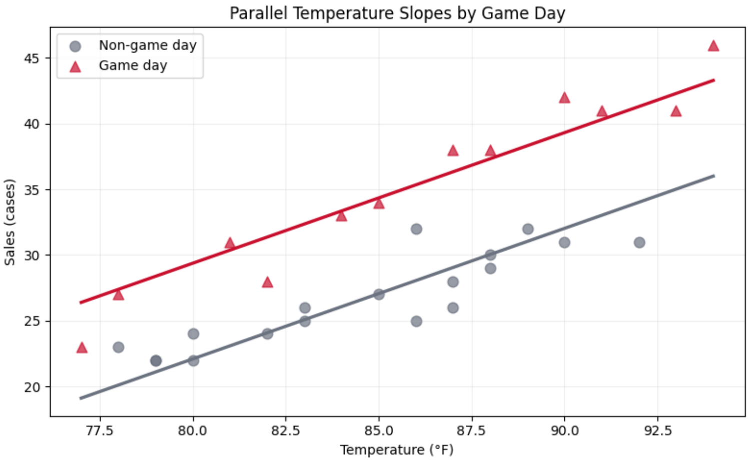
$$\text{Game Day} = \begin{cases} 1 & \text{game day} \\ 0 & \text{non-game day} \end{cases}$$

- Sales (Y); Temperature (X_1); Game Day (X_2)
- How does the linear model look like?

$$\begin{aligned} E(Y|X_1, X_2) &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 \\ &= \begin{cases} \beta_0 + \beta_1 X_1 + \beta_2 & \text{for game day} \\ \beta_0 + \beta_1 X_1 & \text{for non-game day} \end{cases} \end{aligned}$$

- How do we interpret this model?
- Can you draw a graph to illustrate this model?

Parallel Fitted Lines



More than two categories

- Consider another covariate – *Location* (X_3), which has 3 categories: Campus, Downtown, and Stadium.
- How can we write the model?

More than two categories

- Consider another covariate – *Location* (X_3), which has 3 categories: Campus, Downtown, and Stadium.
- How can we write the model?
  Following the same logic, we can write:

$$E(Y|X_1, X_2, X_3) = \begin{cases} \beta_0 + \beta_1 X_1 + \beta_2 X_2 & \text{for Campus} \\ \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{3,1} & \text{for Downtown} \\ \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{3,2} & \text{for Stadium} \end{cases}$$

- Can you see how do we handle the categorical variable *Location* (X_3)?
- How do we draw conclusions in testing the significant of X_3?

Python output

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Sales   R-squared:                       0.689
Model:                            OLS   Adj. R-squared:                  0.653
Method:                 Least Squares   F-statistic:                     19.22
Date:                Wed, 09 Sep 2026   Prob (F-statistic):           8.92e-07
Time:                        17:11:06   Log-Likelihood:                -81.306
No. Observations:                  30   AIC:                             170.6
Df Residuals:                      26   BIC:                             176.2
Df Model:                           3
Covariance Type:            nonrobust
=============================================================================================
                                coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------
Intercept                   -60.0970     13.645     -4.404      0.000     -88.144     -32.050
C(Location)[T.Downtown]       1.5354      1.891      0.812      0.424      -2.351       5.422
C(Location)[T.Stadium]        4.1102      1.791      2.295      0.030       0.428       7.792
Temperature                   1.0374      0.164      6.321      0.000       0.700       1.375
==============================================================================
Omnibus:                        8.510   Durbin-Watson:                   0.801
Prob(Omnibus):                  0.014   Jarque-Bera (JB):                2.199
Skew:                          -0.023   Prob(JB):                        0.333
Kurtosis:                       1.674   Cond. No.                     1.63e+03
==============================================================================
```

Can you write the estimated model and interpret the coefficients?

Interaction

- Consider the same example without Location.

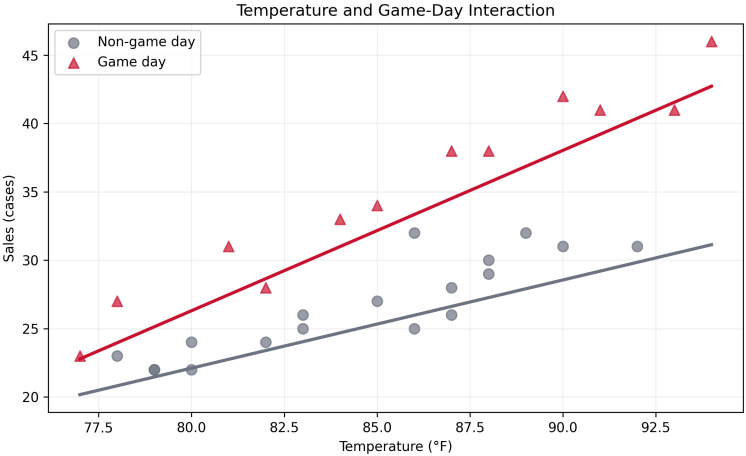
$$E(Y|X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$
$$= \begin{cases} \beta_0 + \beta_1 X_1 + \beta_2 & \text{for game day} \\ \beta_0 + \beta_1 X_1 & \text{for non-game day} \end{cases}$$

- This model says Temperature (X_1) has the same effect regardless it is a game day or not. But what if the effects are different?
- To answer this question, we need to examine the interaction effect

$$E(Y|X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 * X_2)$$

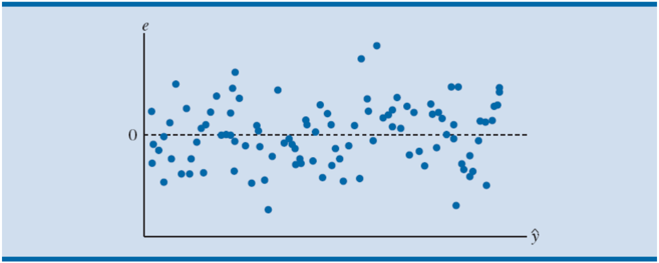
- Can you break down this model with respect to GameDay?
- Can you draw a graph to illustrate this model?

Parallel Fitted Lines



Model Diagnostics

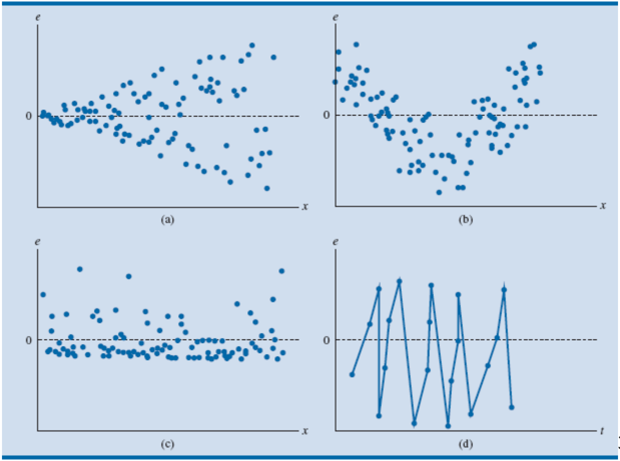
- Model assumptions:
  - Linear relationships between Y and X's
  - The error term $\{\epsilon_i, \ldots, \epsilon_n\} \overset{i.i.d.}{\sim} N(0, \sigma^2)$
    Independent normal distribution; $\mathbb{E}(\epsilon_i) = 0$; $Var(\epsilon_i) =$constant.
- Residual plot (an ideal residual plot looks like this)[2]



[2]source: Camm, et al., *Essentials of Business Analytics*

Residual plot

Which type of assumption is violated?



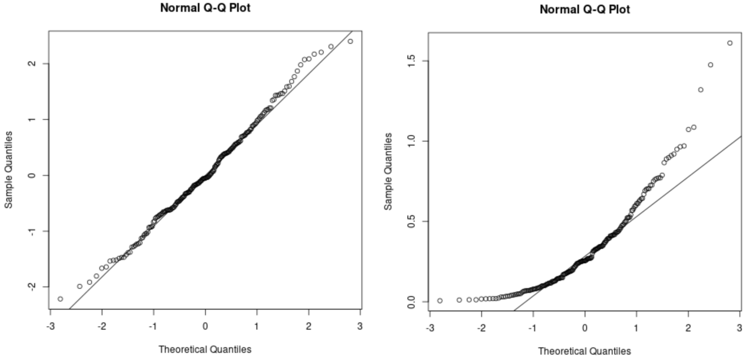
3

[3] source: Camm, et al., *Essentials of Business Analytics*

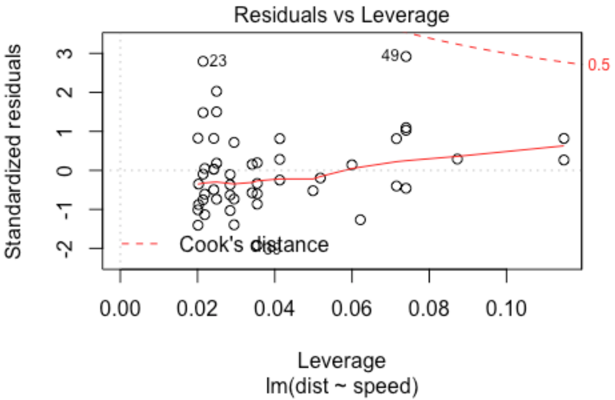
Other Diagnostic Plots

- Normal Quantile-Quantile Plot
  - It plots the standardized residual vs. theoretical quantiles
  - An easy way to visually test the normality assumption
  - If residual follows normal distribution, you should expect all dots lie on the diagonal straight line.
- Residual-Leverage Plot
  - This plot checks if there are any influential points, which could alter your analysis by excluding them
  - The points that lie outside the dashed line, Cook's distance, are considered as influential points

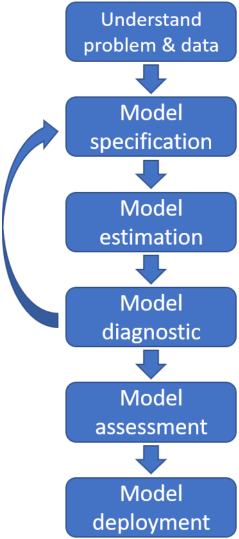
Normal Q-Q plot



Residual leverage plot



Model Building Process



- Describe the problem by stating the goal and purpose of statistical analysis.
- Given the dataset, conduct exploratory data analysis to maximize the knowledge of information carried by the data.

- Identify relevant variables and do necessary data cleaning and transformation. These include missing data imputation, variable recoding and transformation, etc.
- Identify appropriate statistical models to achieve the purpose of analysis.

- Use software to estimate the model. This process may involve debugging.
- Understand the main components of the output.
- Statistical inference needs to be conducted.

- Check if there is any model assumption violation. (For more ML oriented methods such as neural networks, this step is often skipped due to less assumptions.)
- If violation is identified, go back and re-specify the model. For example, if residual plots show quadratic pattern, then the quadratic term of one or more predictors needs to be included. Graphical inspections are needed in this procedure.

- Typically, various models are estimated, including a certain type of model with different sets of variables (determined by variable selection), and different types of model such as linear regression, regression tree, and neural networks.
- Compare all estimated models in terms of certain criteria, such as cross-validation error, AIC and BIC, etc.

- At the end of analysis, a final model needs to be determined for deployment.
- Depending on different scenarios, the final model may be a set of models that include several different models.
- Inference and model interpretation should be part of the deployment.