Linear Regression [1]

Shaobo Li

University of Kansas

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

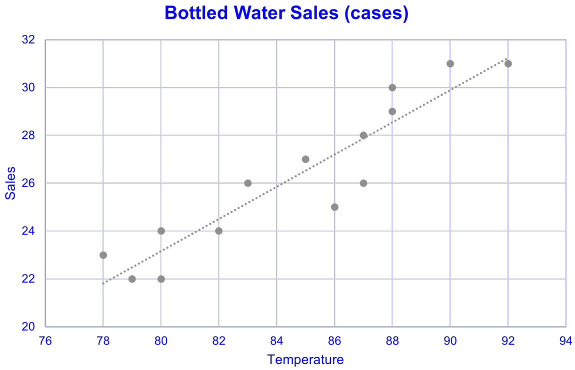
Linear regression – a fundamental learning algorithm

- 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?

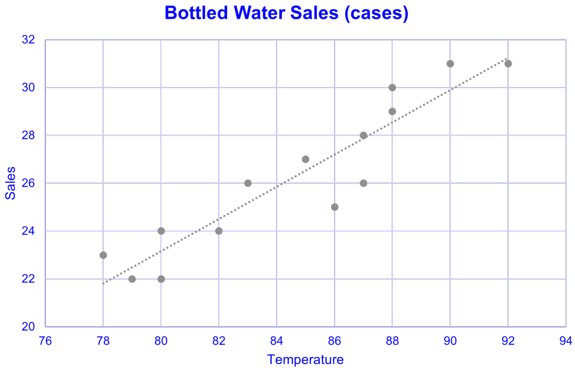
- Is there an association between X and Y?
- If yes, how strong is this association?
- Is this association linear?
- If there are multiple X's (X_1, X_2 and X_3), which of them are related to Y and which are not?
- Can we predict the value of Y for any given X?
- How accurate is such prediction?

An example – bottled water sales



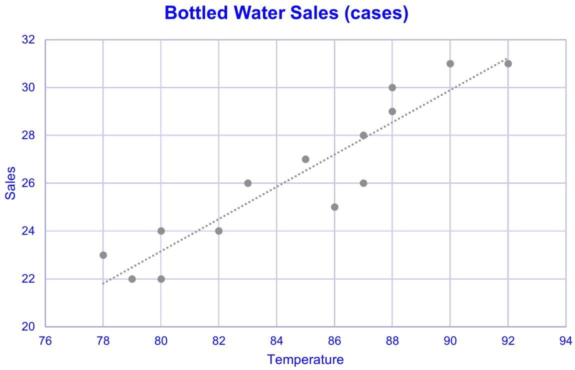
- Is there an association between *Temperature* and *Sales*?

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 model alway 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

$$Sales \approx -30.70 + 0.67 \times Temperature$$

- This is the straight line we saw before.

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 between the "model" and 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

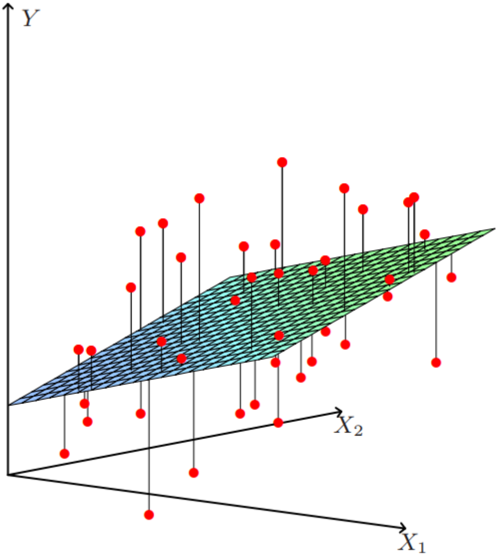
Model estimation

- 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)

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

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

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

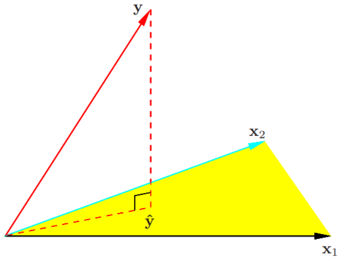
- This is called *normal equation*.
- Exercise: derive the closed form of the solution

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

- 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 = sales)
> summary(model1)

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

Residuals:
    Min      1Q  Median      3Q     Max
-2.1994 -0.5016  0.2908  0.8350  1.4542

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -30.69720    6.38033  -4.811 0.000425 ***
Temperature   0.67322    0.07529   8.942 1.18e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.184 on 12 degrees of freedom
Multiple R-squared:  0.8695,	Adjusted R-squared:  0.8586
F-statistic: 79.96 on 1 and 12 DF,  p-value: 1.182e-06
```

Bootstrap

- Resampling method
- A powerful tool to quantify uncertainty
  - standard error
  - confidence interval
- Random sampling with replacement
- The general procedure:
  - fit a model **B** times based on **B** bootstrap samples
  - store all the parameter estimates
  - calculate standard error and confidence interval

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?)

Model Assessment – AIC, BIC, Cp, CV

- Akaike information criterion (AIC), the smaller the better

$$AIC = -2\log(\hat{L}) + 2p$$

- Bayesian information criteria (BIC), the smaller the better

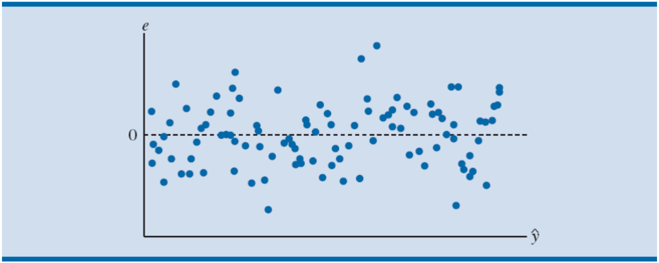
$$BIC = -2\log(\hat{L}) + \log(n)p$$

  where $\hat{L}$ is estimated likelihood function
- Mellow's C_p is the same as AIC for linear regression
- Cross-validation error (to assess prediction accuracy)
- These metrics are important for model selection

Model Diagnostics

- Model assumptions:
  - Linear relationships between Y and X's
  - The error term $\{\epsilon_i, \ldots, \epsilon_n\} \overset{i.i.d.}{\sim} N(0, \sigma^2)$

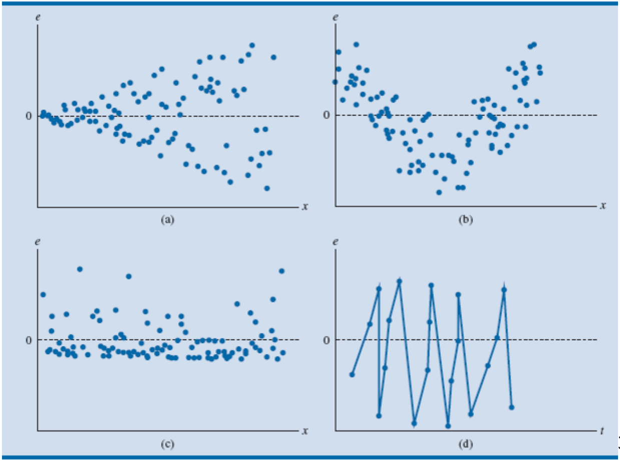
    Independent normal distribution; $\mathbb{E}(\epsilon_i) = 0$; $Var(\epsilon_i) =$constant.
- Residual plot (an ideal residual plot looks like this)[2]



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

Residual plot

Which type of assumption is violated?



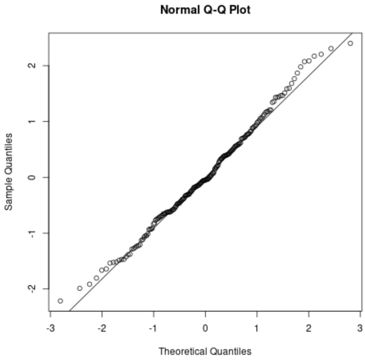
[3]

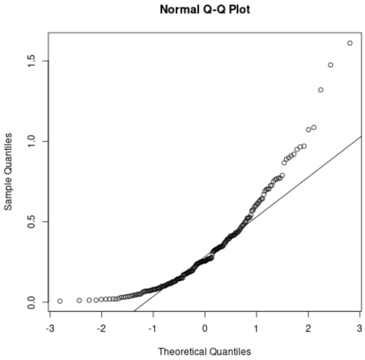
[3] source: Camm, et al., *Essentials of Business Analytics*

Other Diagnostic Plots

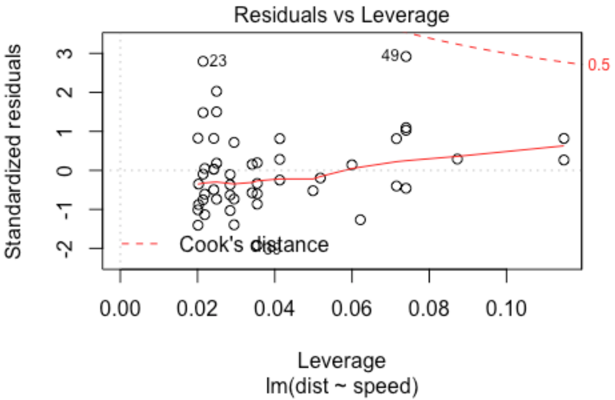
- Normal Quantile-Quantile Plot
  - It plots the standardized residual vs. theoretical quantiles
  - An easy way to visually test the normality assumption
  - If residual follows normal distribution, you should expect all dots lie on the diagonal straight line.
- Residual-Leverage Plot
  - This plot 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

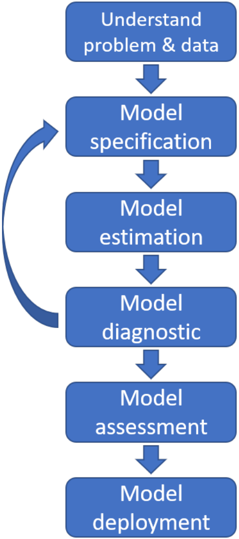




Residual leverage plot



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.