Linear Models: Application to Marketing

Aramayis Dallakyan[1] [2]

[1] Agribusiness teaching Center
Armenia

2018

[2] All errors are my own. armopost@yahoo.com

Outline

1. Introduction

2. Linear Models
   - Overview of Least Squares Estimate
   - Estimating the population slope and intercept

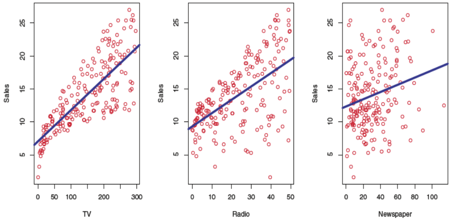
Introduction

Introduction

- In this lecture note we will focus on tools and techniques for building valid regression analysis for real-world data, in particular we will concentrate on Linear Regression :
- We shall see that a key step in any regression analysis is assessing the validity of the given model. When weaknesses in the model are identified we need to address them as correct as possible. An important thing to remember is that
- **It makes sense to base conclusions only on valid models**

Motivation

- To motivate our study of statistical learning, we begin with the simple example.
- Suppose you are marketing consultant hired by client to provide advice on how to improve sales of a particular product.
- The data you have is following



Our main goal is to determine is there a relationship between advertising and sales??

Introduction

Generally, suppose that we observe a quantitative response Y and p different predictors, $X_1, \ldots, X_p$. We assume that there some relationship between Y and $X = (X_1, \ldots, X_p)$, which can be written in the general form

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

, where usually f is unknown and in practice our job is to approximate f as close as possible.

Question: Why we need to estimate f ?

Introduction

There are two main reasons that we may wish to estmate f.

- **Prediction:** our job is to model $\hat{Y} = \hat{f}(X)$ such that we predict f as accurate as possible. This is very importent topic, but we are not going to cover it.

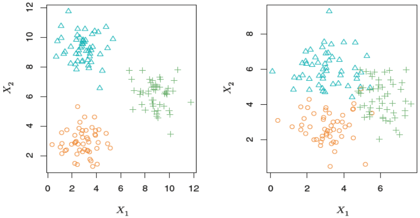
$$E(Y - \hat{Y}) = \underbrace{[f(X) - \hat{f}(X)]^2}_{reducible} + \underbrace{Var(\epsilon)}_{irreducible}$$

- **Inference:** we are often interested in understanding the way that Y is affected as $X_1, \ldots, X_p$ change.

Supervised vs Unsupervised

Most statistical problems fall into one of two categories:**Supervised** and **Unsupervised**.

- In supervised learning, for each observation of the predictor measurement $x_i, i = 1, \dots, n$ there is an associated response measurement y_i. For example advertising and sales.
- In contrast, supervised learning describes the somewhat more challenging situation in which for every observation $i = 1, dots, n$, we observe a vector of measurements x_i, but no associated response y_i.



Introduction

Our interest is in answering the following questions:

1. How to choose f to minimize "reducible" error?
2. Which predictors are associated with the response? That is identify the few important predictors.
3. What is the relationship between the reposnse and each predictor?
4. Can the relationship between Y and each predictor be adequately summarized using a linear equation, or is the relationship more complicated.

Linear Models

Linear Models

In linear models we assume that the functional form, or shape of f is linear. That is

$$f(x) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- Linear models used in marketing to explore relationship between outcome of interest and other variables.
- A common application in survey analysis is to model satisfaction with a product in relation to specific elements of the product and its delivery; this is called **"satisfaction drivers analysis."**
- Linear models are also used to understand how price and advertising are related to sales, and this is called **"marketing mix modeling."**

Linear Models

- In this class, we illustrate linear modeling with a satisfaction drivers analysis using survey data for customers who have visited an amusement park.
- In the survey, respondents report their levels of satisfaction with different aspects of their experience, and their overall satisfaction.
- Marketers frequently use this type of data to figure out what aspects of the experience drive overall satisfaction, asking questions such as,
- Are people who are more satisfied with the rides also more satisfied with their experience overall?" If the answer to this question is "no," then the company will know to invest in improving other aspects of the experience.

Least Squares

1. Regression analysis is a method for investigating the functional relationship among variables.
2. Plots will be an important tool for both building regression models and assessing their validity. We shall see that deciding what to plot and how each plot should be interpreted will be a major challenge.
3. Let start by reviewing simple linear models involving modeling the relationship between two variables.

Least Squares

- In particular we consider problem involving modeling the relationship between two variables as a straight line, that is, when Y is modeled as a linear function of X .
- Suppose we have data :

  $$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

  where x_1 denotes the first value of the so-called X -variable and y_1 denotes the first value of the so-called Y -variable.
- **Question:** What are the usual names for X and Y?

Least Squares

- Simple linear regression is typically used to model the relationship between two variables Y and X so that given a specific value of X , that is, $X = x$, we can predict the value of Y .

- Mathematically, the regression of a random variable Y on a random variable X is

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

, the expected value of Y when X takes the specific value x .

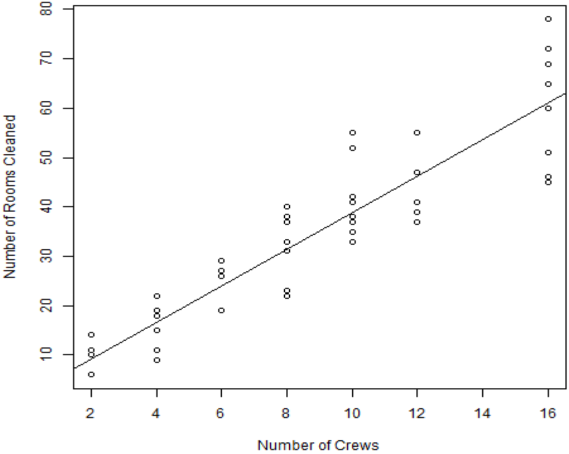
Lease Square

- For example, you work for the famous hotel and want to model realtionship $X =$ Number of Crews and $Y =$, Number of Rooms Cleaned, then the regression of Y on X represents the mean (or average) cleaned room on a given number of crew.
- So the regression of Y on X is linear if

$$E(Y|X = x) = \beta_0 + \beta_1 x$$

, where β_0 and β_1 are the intercept and the slope of a specific straight line, respectively.

Example of Least Squares



Least Squares

Suppose that $Y_1, Y_2, \ldots, Y_n$ are independent and identical (explain i.i.d?) realizations of the random variable Y (explain r.v ?) that are observed at the values $x_1, x_2, \ldots, x_n$ of a random variable X . If the regression of Y on X is linear, then for $i = 1, 2, \ldots, n$

$$Y_i = E(Y|X = x) + e_i = \beta_0 + \beta_1 x + e_i$$

Question: Explain what is e_i here and why we need error. Also what other assumption usually we make?

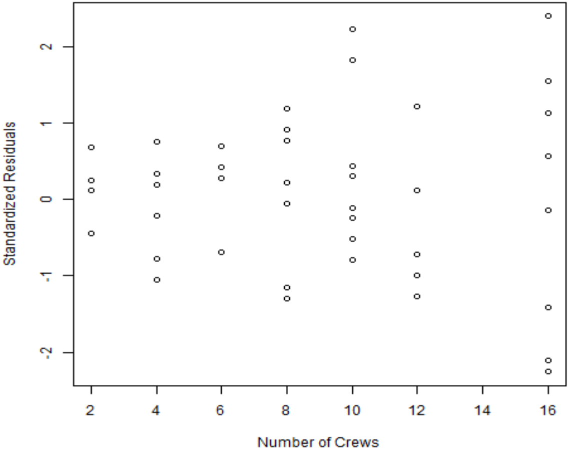
Least Squares

- The random error term captures all unexplained variation .
- Thus, the random error term does not depend on x , nor does it contain any information about Y . For now let assume

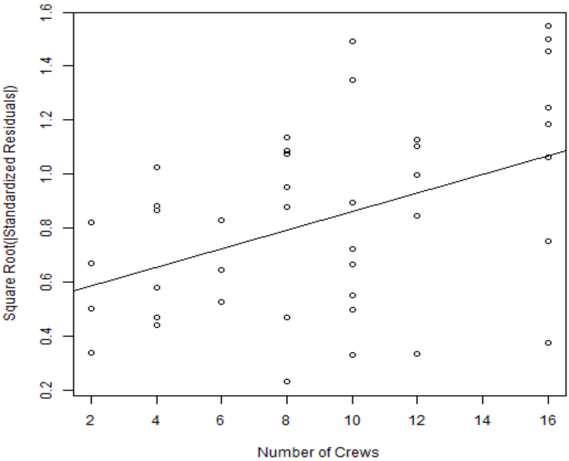
$$Var(Y|X = x) = \sigma^2$$

- **Question:** What is the name of the model where variance is not constant and what would happen if we ignore the problem.

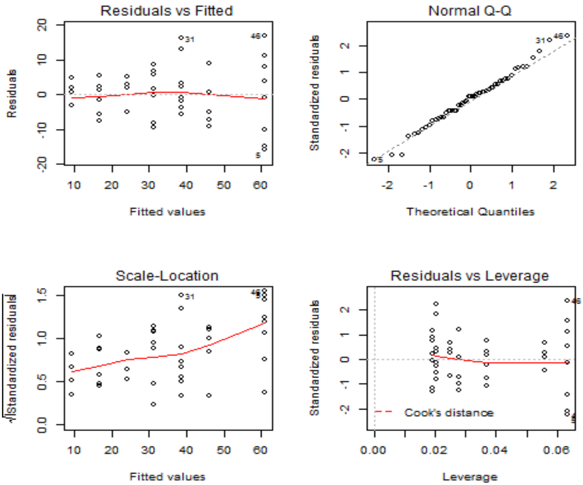
Hotel Example cnt.



Hotel Example cnt.



Hotel Example cnt.



OLS Estimation

- Ideally we want correct values for β_0 and β_1. Unfortunately β_0 and β_1 are always unknown , since they represent the true population and **we never know the truth.**
- Thus, our task is use a sample of data instead of the whole population. That is use the given data to estimate the slope and the intercept.
- This can be achieved by finding the equation of the line which **"best" fits** our data, that is, choose b_0 and b_1 such that $\hat{y}_i = b_0 + b_1 x$ is as **"close"** as possible to y_i .

OLS Estimation

In practice, we wish to minimize the difference between the actual value of $y(y_i)$ and the predicted value of $y(\hat{y}_i)$. This difference is called the residual, $\hat{e}_i$, that is,

$$\hat{e}_i = y_i - \hat{y}_i$$

.

A very popular method of choosing b_0 and b_1 is called the method of least squares. (Explain LS?)

OLS Estimation

As the name suggests b_0 and b_1 are chosen to minimize the sum of squared residuals RSS.

$$RSS = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

For RSS to be a minimum with respect to b_0 and b_1 we require (what is the name of this requirement?)

$$\frac{\partial RSS}{\partial b_0} = -2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) = 0$$

and

$$\frac{\partial RSS}{\partial b_1} = -2 \sum_{i=1}^{n} x_i (y_i - b_0 - b_1 x_i) = 0$$

OLS Estimation

These last two equations are called the **normal equations** . Solving these equations for b_0 and b_1 gives the so-called least squares estimates of the intercept (VFY)

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

and the slope

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Also, we can use the residuals $\hat{e}_i$ to estimate the true σ^2. In fact it can be shown that

$$S^2 = \frac{RSS}{n-2} = \frac{1}{n-2} \sum \hat{e}_i{}^2$$

Question: What we can say about $\sum \hat{e}_i$ and why we have 2 in denominator?

OLS Estimation

When we have more than one explanatory variable we deal with **multiple linear regression**.

$$E(Y|X_1 = x_1, \ldots, X_p = x_p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

Thus,

$$Y_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

In matrix notation the problem becomes

$$RSS(\beta) = (\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta)$$

after OLS one can show that

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

It is easy to show that

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

$$Var(\hat{\beta}|\mathbf{X}) = \sigma^2(\mathbf{X^T X})$$

Numerical Results

Assuming that errors are normally distributed $e_i \sim N(0, \sigma^2)$, it can be shown that for $i = 0, 1 \ldots, p$

$$T_i = \frac{\hat{\beta}_i - \beta_i}{se(\hat{\beta})} \sim t_{n-p-1}$$

Thus we can do hypothesis testing such as

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

Numerical Results

When we reject null hypothesis? How we test whether there is a linear association between Y and $X_1, \dots, X_p$? If

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + e$$

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

$$H_A : \text{at least some of the } \beta_i \neq 0$$

If $SSreg = \sum(\hat{y}_i - \bar{y})^2$ than we can test the hypothesis by

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

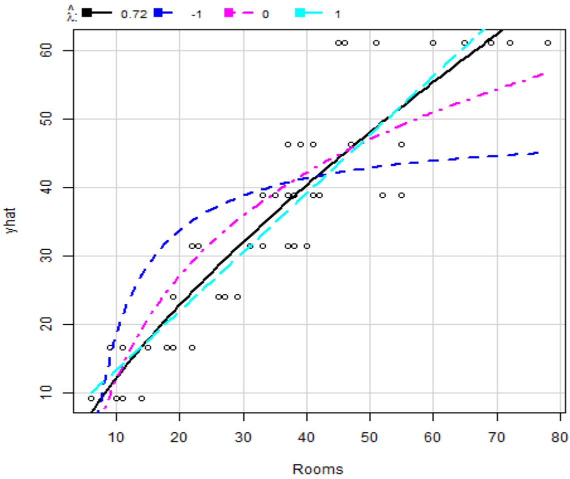
One last cool thing you may want to do is to check the linearity assumption of the whole function $f(x)$.
Suppose the true regression model between Y and X is given by

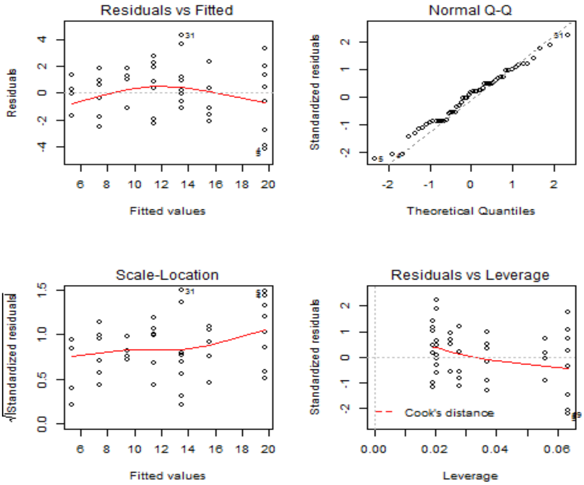
$$Y = g(\beta_0 + \beta_1 x + \epsilon)$$

, where g is unknown. The the inverse of g is

$$g^{-1}(Y) = \beta_0 + \beta_1 x + \epsilon$$

. Thus if we knew β_0 and β_1 we can discover the shape of g^{-1} by plotting Y on the horizontal axis and $\beta_0 + \beta_1 x$ on the vertical axis.
In practice since β's are unknown, given that some theoretical assumptions are satisfied, g^{-1} can be estimated using fitted values $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ Such plot is called **inverse response plot**


$\hat{\lambda}$: 0.72
-1
0
1
yhat
10
20
30
40
50
60
10
20
30
40
50
60
70
80
Rooms


Residuals vs Fitted
Residuals
4
2
0
-2
-4
31
5
8
6
8
10
12
14
16
18
20
Fitted values
Normal Q-Q
Standardized residuals
2
1
0
-1
-2
31
5
-2
-1
0
1
2
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
1.5
1.0
0.5
0.0
31
5
8
6
8
10
12
14
16
18
20
Fitted values
Residuals vs Leverage
Standardized residuals
2
1
0
-1
-2
Cook's distance
0.00
0.02
0.04
0.06
Leverage

	Dependent variable:	
	Rooms	Rooms_trsf
	(1)	(2)
Crews	3.701***	1.027***
	(0.212)	(0.057)
Constant	1.785	3.278***
	(2.096)	(0.561)
Observations	53	53
R^2	0.857	0.865
Adjusted R^2	0.854	0.863
Residual Std. Error (df = 51)	7.336	1.964
F Statistic (df = 1; 51)	305.275***	327.928***

Now we can compare two models based on three more metrics such as **adjusted-R^2**,**AIC** and **BIC**.The rule of thumb for **AIC** and **BIC** criteria are smaller the value better the model.

Table

	m1	m2
R_sq	0.854	0.863
AIC	213.690	73.978
BIC	217.141	77.429

Numerical Results

Next we are going to analyze satisfaction drivers using survey data for customers who have visited an amusement park. The main points we are going to spent time on are:

1. Determine whether the proposed regression model is a valid model .(plots of standardized residuals) .
2. Determine which (if any) of the data points have x -values that have an unusually large effect on the estimated regression model
3. Determine which (if any) of the data points are outliers , that is, points which do not follow the pattern.
4. Examine whether the assumption of constant variance of the errors is reasonable.
5. For small sample examine whether the assumption that the errors are normally distributed is reasonable.

Numerical Results

Amusemant park has 8 observed predictors. That is we have $2^8 = 256$ possible models.

1. $M_0 : y = \beta_0 + \beta_1 x_1 + \epsilon$
2. $M_2 : y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \epsilon$
3. $\vdots \vdots \vdots \vdots \vdots$
4. $M_7 : y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_8 x_8 + \epsilon$
5. $M_8 : y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_8 x_8 + \beta_9 x_1 x_2 + \epsilon$
6. $\vdots \vdots \vdots \vdots \vdots$