IEOR165 Discussion
Week 2

Sheng Liu

University of California, Berkeley

Jan 29, 2016

Outline

Office Hour, Homework and Project

- Office hour: 4-5 pm on Thursdays and Fridays, 1174B Etch Hall.
- Homework: the first homework will be published next Tuesday.
- Project: more details will be released.

Homework will not require coding but project will do!
You can use Matlab, `R`, Python...
My preference: `R`

- Download it from `http://cran.r-project.org`
- `Rstudio`: a popular integrated development environment (IDE), `https://www.rstudio.com/products/rstudio/download/`
- The online documentation and installation routines are comprehensive.

Linear Regression: Purposes

- Find relationships between variables, make predictions.
- Linear regression is simple but powerful.
- Many advanced statistical learning approaches can be seen as generalizations or extensions of linear regression models.

The Model:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r + \epsilon \quad \text{(Multiple Regression)}$$

$$y = \beta_0 + \beta_1 x + \epsilon \quad \text{(Simple Regression)}$$

- y : response, output, dependent variable
- x : predictors, inputs, independent variables, features
- β : parameters, coefficients (intercept and slope)

Linear Regression: Linearity

Berkeley UNIVERSITY OF CALIFORNIA

One of the key assumptions in linear regression model is the linearity assumption. This assumption is actually not that restrictive:

An example[1]: suppose we have a data set that includes

- Q: Demand
- P: Price
- M: Consumer Income
- P_R: Price of Some Related Good R

Under some economic assumptions, we have the following demand function:

$$Q = aP^b M^c P_R^d$$

How to utilize linear regression model to estimate a, b, c, d?

[1] http://highered.mheducation.com/sites/0073402818/student_view0/chapter7/index.html

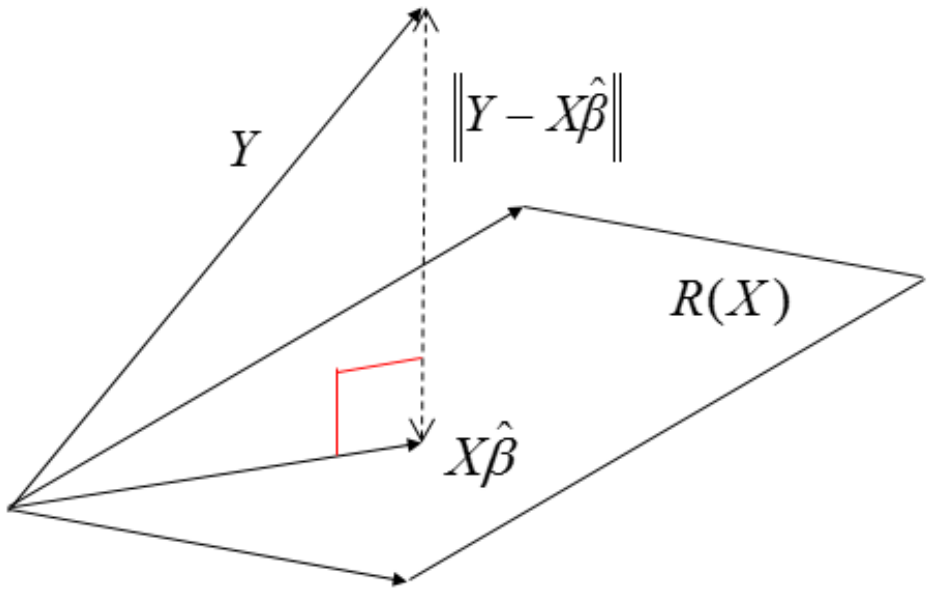
Linear Regression: Estimates

- In practice, we will be given a data sample with n observations: $(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)$. We want to estimate the parameters from the data set: method of least squares. The key idea: find parameters that minimize the sum of squared errors $\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$
- The geometric interpretation:

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \quad \hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix}$$

We know $rank(X) = 2$, which means that $\mathcal{R}(X)$ is 2-dimensional subspace in $\mathbb{R}^n$, i.e. a 2-dimensional plane. So $\hat{Y} = X\hat{\beta}$ is in this plane, and $\min ||Y - \hat{Y}||_2^2$ will be the squared distance from Y to this plane.

Linear Regression: Estimates



The minimum is reached when $Y - \hat{Y}$ is orthogonal to $\mathcal{R}(X)$, i.e. $\hat{Y}$ is the projection of Y onto the 2-dimensional plane. Thus we have

$$X^\top (Y - \hat{Y}) = X^\top (Y - X\hat{\beta}) = 0$$

$$\Rightarrow X^\top Y = X^\top X\hat{\beta} \quad \Rightarrow \hat{\beta} = (X^\top X)^{-1} X^\top Y$$

Linear Regression: Example

(ISLR[2]) Suppose that we are statistical consultants hired by a client to provide advice on how to improve sales of a particular product. The Advertising data set consists of the sales of that product in 200 different markets, along with advertising budgets for the product in each of those markets for three different media: TV, radio, and newspaper. It is not possible for our client to directly increase sales of the product. On the other hand, they can control the advertising expenditure in each of the three media. Therefore, if we determine that there is an association between advertising and sales, then we can instruct our client to adjust advertising budgets, thereby indirectly increasing sales. In other words, our goal is to develop an accurate model that can be used to predict sales on the basis of the three media budgets.

1, Identify the response and the predictors.
2, Establish the linear model.
3, If we are only interested in the relationship between TV budgets and sales, what will be the linear model?

[2] James, Gareth, et al. An introduction to statistical learning. New York: springer, 2013.

Linear Regression: Estimates

62 3. Linear Regression

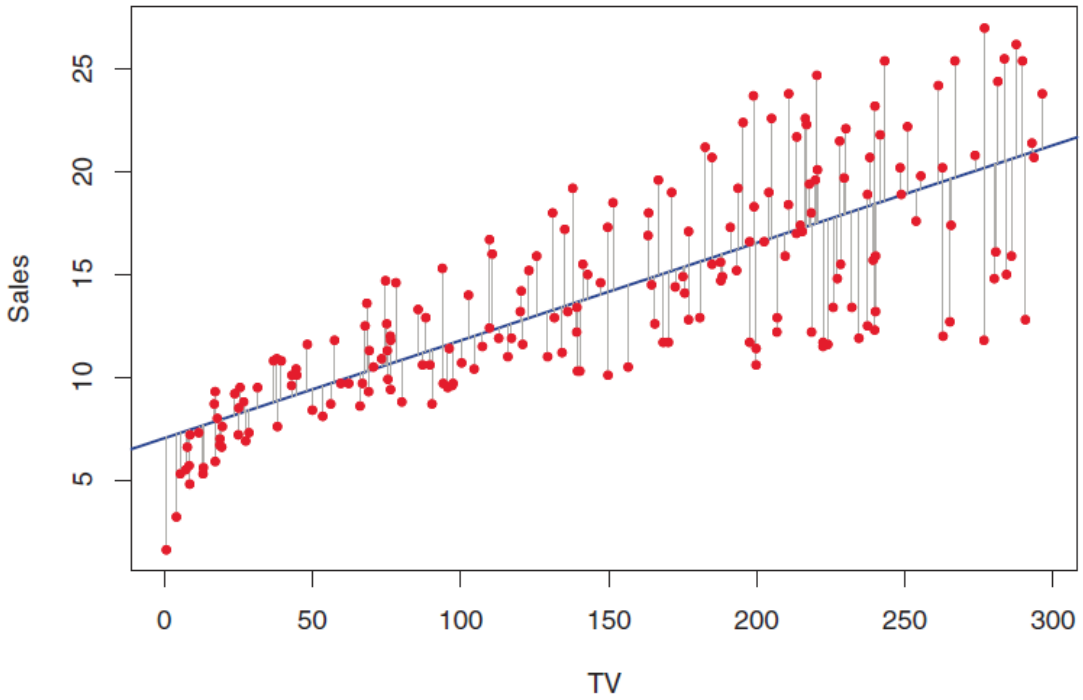


FIGURE 3.1. *For the* Advertising *data, the least squares fit for the regression of* sales *onto* TV *is shown. The fit is found by minimizing the sum of squared errors. Each grey line segment represents an error, and the fit makes a compromise by averaging their squares. In this case a linear fit captures the essence of the relationship, although it is somewhat deficient in the left of the plot.*

Suppose we get $\hat{\beta}_1 = 0.0475$, then how will you predict the sales if we spend an additional \$1000 on TV advertising?

Linear Regression: Qualitative Variables

Which of the following variables are qualitative variables (categorical variables)?

- Age
- Income
- Gender
- Nationality
- Favorite Band

Incorporate qualitative variables in linear regression:

1. If we have a data set including *gender* as a variable, how do you incorporate it in the linear regression model?
2. If we have a data set including *favorite band* $B \in \{\text{The Beatles, Coldplay, Nirvana, Led Zeppelin}\}$ as a variable, how do you regress the *Income* on the *favorite band*?

Linear Regression in R

The Ball Trajectory Example:

```
x=c(0.56, 0.61, 0.12, 0.25, 0.72, 0.85, 0.38, 0.90, 0.75, 0.27)

y=c(0.25, 0.22, 0.10, 0.22, 0.25, 0.10, 0.18, 0.11, 0.21, 0.16)

lm.fit=lm(y~x)

lm.fit

##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##   0.1795941    0.0007502
```

Linear Regression in R

```
summary(lm.fit)

##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.08023 -0.05765  0.01498  0.04015  0.06999
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.1795941  0.0464384   3.867  0.00476 **
## x           0.0007502  0.0774695   0.010  0.99251
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06325 on 8 degrees of freedom
## Multiple R-squared:  1.172e-05,Adjusted R-squared:  -0.125
## F-statistic: 9.378e-05 on 1 and 8 DF,  p-value: 0.9925
```

Linear Regression in R

Berkeley UNIVERSITY OF CALIFORNIA

Residuals Plot:

```
plot(x,residuals(lm.fit),col="blue")
```