CSCI 3022

intro to data science with probability & statistics

November 16, 2018

Multiple Linear Regression

Department of Computer Science
UNIVERSITY OF COLORADO **BOULDER**

Dan Larremore

Stuff & Things

- HW6 due Friday after break.
- Arkaive it up!

Last time on CSCI3022: SLR

- Given data, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ fit a simple linear regression of the form

$$Y_i = \alpha + \beta x_i + \epsilon_i \qquad \epsilon_i \sim N(0, \sigma^2)$$

- Compute estimates of the intercept and slope parameters by minimizing:

$$SSE = \sum_{i=1}^{n} [y_i - (\alpha + \beta x_i)]^2$$

- The least-squares estimates of the parameters are:

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

Last time on CSCI 3022:

- We can perform inference on slope to determine if relationship is significant.

$$\hat{\sigma}^2 = \frac{SSE}{n-2} \qquad se(\hat{\beta}) = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \qquad CI : \hat{\beta} \pm t_{\alpha/2,n-2} \times se(\hat{\beta})$$

- We can use the Coefficient of Determination to evaluate goodness-of-fit of SLR model

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad SSE = \sum_{i=1}^{n}[y_i - (\alpha + \beta x_i)]^2 \qquad R^2 = 1 - \frac{SSE}{SST}$$

- If R^2 is close to 1 then the model fits the data relatively well.

Regression with Multiple Features

In most practical applications there are multiple features or predictors that potentially have an effect on the response.

Example: Suppose that y represents the sale price of a house. Reasonable features associated with sale price might be:

- x_1 : the interior size of the house
- x_2 : the size of the lot on which the house sits
- x_3 : the number of bedrooms in the house
- x_4 : the number of bathrooms in the house
- x_5 : the age of the house

SLR: give me x
I tell you y

MLR: give me $x_1, x_2, x_3 \ldots$
I tell you y.

Regression with Multiple Features

Questions we would like to answer in the next few classes:

- Is at least one of the features useful in predicting the response?
- Do all of the features help to explain the response, or is it just a subset?
- How well does the model fit the data?
- Given a set of predictor values, what response should we predict, and how accurate is our prediction?

We will look at these questions over the course of the week, but first let's do a little exploration of a multiple feature data set and remind ourselves about SLR

Advertising Budget Example

- Get in groups (pairs at least!), get out your laptops, and open the Lecture 22 In-Class Notebook
- **Example**: Data is provided about the sales of a particular product in 200 different markets, along with advertising budgets for each market for three different media types: TV, Radio, and Newspaper.
- The sales response is given in thousands of units, and each of the advertising budget features are given in thousands of dollars.
- We will begin by fitting individual SLR models with the advertising budget as the feature and the sales as a response.

Multiple Linear Regression

$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}$

- We've seen from the Advertising example that SLR analysis has indicated that there is a significant relationship between each of the media types: TV, Radio, and Newspaper on the sales of the product.
- But individual SLR models only show the effect of each media type in a vacuum. To get a clearer picture of what's going on, we want to consider the effect of all three advertising types on sales simultaneously
- This is where Multiple Linear Regression (MLR) comes in
- **Def**: In MLR, the data is assumed to come from a model of the form:

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

(β_0: intercept)

X_1 = news
X_2 = radio
X_3 = TV
$p = 3$

Multiple Linear Regression

- This means that for each of n data points $(x_{i1}, x_{i2}, \ldots, x_{ip}, y_i)$ we assume

response intercept feature values

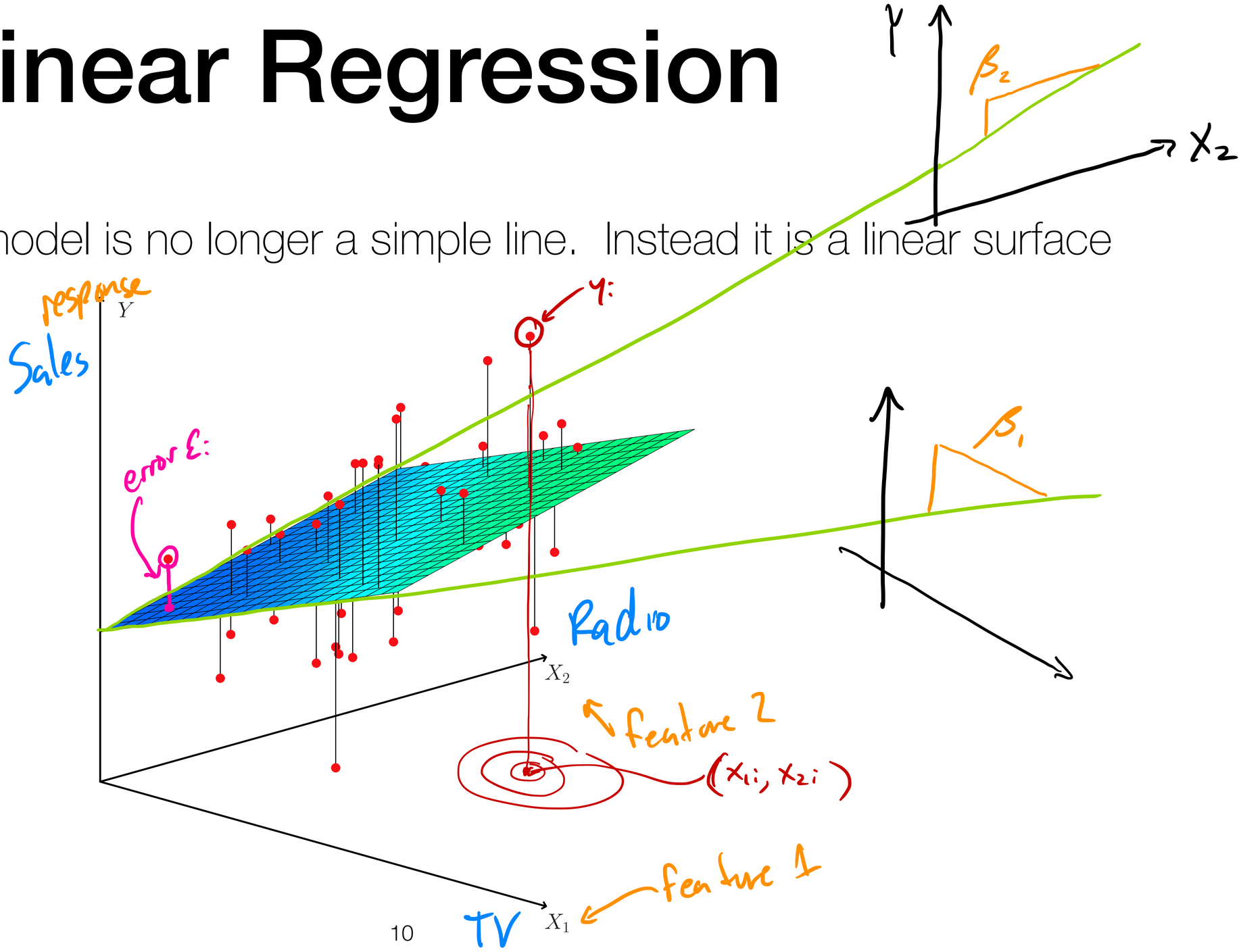
$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$$

"slopes" noise!

- We make similar assumptions as in the case of SLR:
  - ϵ_i are independent of each other
  - $\epsilon_i \sim N(0, \sigma^2)$

Multiple Linear Regression

- Note that our model is no longer a simple line. Instead it is a linear surface



Multiple Linear Regression

intercept slopes.

- The interpretation of the model parameters are similar to that of SLR

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$$

- Parameter β_k is the expected change in the response associated with a 1-unit change in the value of x_k while all other features are held fixed.

- **House Sale Price Example**:

$y = 100 + 3x_1 + 2x_2 + 2.5x_3$

If I have a house w/ same lot size (x_2) and same sq. ft. (x_1), then if I ↑ the # of cool dogs in the neighborhood (x_3) by 1 cool dog, then price(y) ↑ 2.5.

Estimating the MLR parameters

- Just as in the case of SLR, we have no hope of discovering the true model parameters, and so have to estimate them from the data. Our estimated model will be

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

- As before, we will determine the estimated parameters by minimizing the sum of squared errors:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_p x_{pi} \right) \right)^2$$

- The SSE is again interpreted as the measure of how much variation is left in the data that cannot be explained by the model.

- Note: Without linear algebra, it is difficult to write down a closed-form expression for the parameter estimates. For now we will simply see how we can find them in Python. Later we'll see how to estimate parameters using the method of **Stochastic Gradient Descent**.

Advertising Budget Example

- Group back up! Let's see how we can find an MLR model for the Advertising data…

Advertising Budget Example

- OK. We've determined that the MLR model for the advertising data is:

$$\texttt{sales} = 2.94 + 0.046 \times \texttt{TV} + 0.189 \times \texttt{radio} - 0.001 \times \texttt{news}$$

- **Question**: Why did our SLR models indicate a positive relationship between newspaper advertising and product sales, but our MLR model did not?

With multiple features, we must consider the relationship between the features themselves.

If 2 features have a strong rel'ship w/ each other, then it's possible only one has a relationship w/ response.

A Correlation Parable

- **Example**: A simple linear regression analysis of **shark attacks vs ice cream sales** at a Southern California beach indicates that there is a strong relationship between the two.
- **Question**: Do you think that this relationship is real?

?
=

A Correlation Parable

- **Example**: A simple linear regression analysis of **shark attacks vs ice cream sales** at a Southern California beach indicates that there is a strong relationship between the two.
- **Question**: Do you think that this relationship is real?
- **Answer**: Probably not. Higher temps cause more people to head to the beach, increasing the chance of shark attacks. And, higher temps cause more people to buy ice cream.
- If we ran a MLR analysis with shark attacks as the response and temperature and ice cream sales as features, our model would show the strong relationship between temperature and shark attacks, and an insignificant relationship between shark attacks and ice cream sales!
- In such an analysis, we say that when we **adjust** or **control** for temperature, the relationship between ice cream sales and shark attacks disappears.

Advertising Budget Example

beautiful

- **Question**: Based on our ~~rather absurd~~ shark attack example, can you explain why newspaper spending became less significant in our MLR of product sales?

$$\texttt{sales} = 2.94 + 0.046 \times \texttt{TV} + 0.189 \times \texttt{radio} - 0.001 \times \texttt{news}$$

Newspaper $ was a surrogate for one of our other features.

When we control for radio and TV, relationship between news + sales disappears!