Multiple Linear Regression

Administrivia

- **Homework** due Friday.
- Good **Milestones**:
    - Problems 1-3 last week
    - Problems 4-6 this week

Previously on CSCI 3022

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

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

Estimates of the intercept and slope parameters are estimated by minimizing

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

The least-squares estimate of the parameters are

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

Previously on CSCI 3022

We can perform inference on slope parameter 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

Regression with Multiple Features

Questions we would like to answer this week:

- 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, 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

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

$\epsilon \sim N(0, \sigma^2)$

Def: In MLR, the data is assumed to come from a model of the form

P FEATURES $(p = 3)$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

Multiple Linear Regression

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

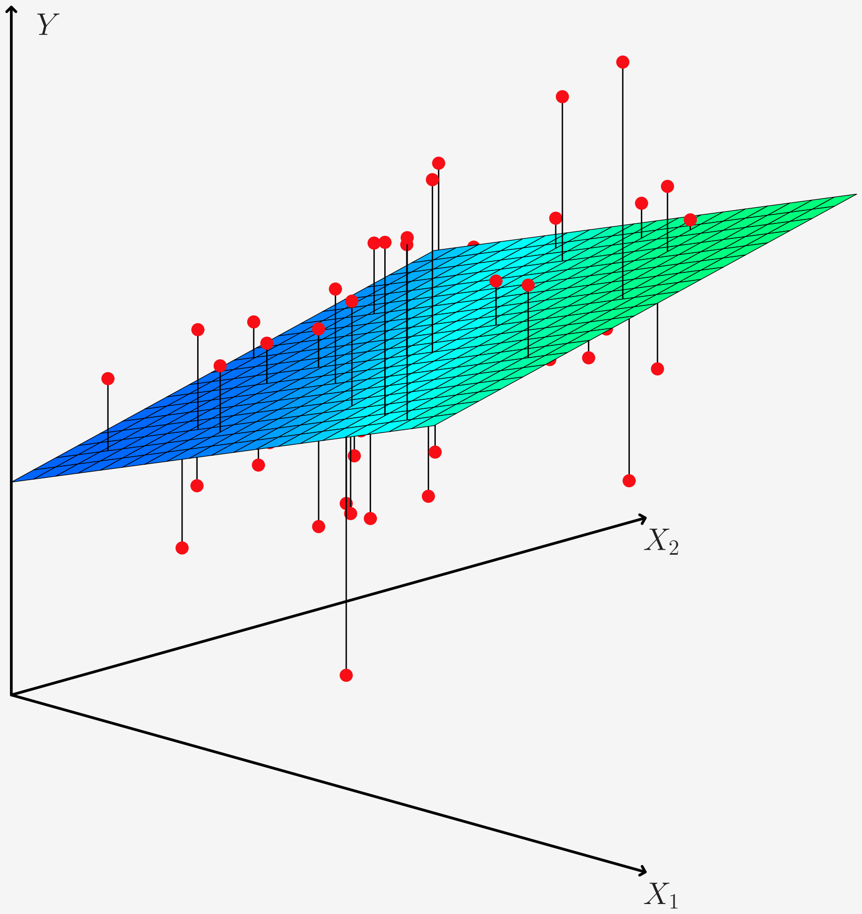
$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$$

We make similar **assumptions** as in the case of SLR:

- EACH ϵ_i ARE INDEPENDT
- $\epsilon_i \sim N(0, \sigma^2)$

Multiple Linear Regression

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



Multiple Linear Regression

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.**

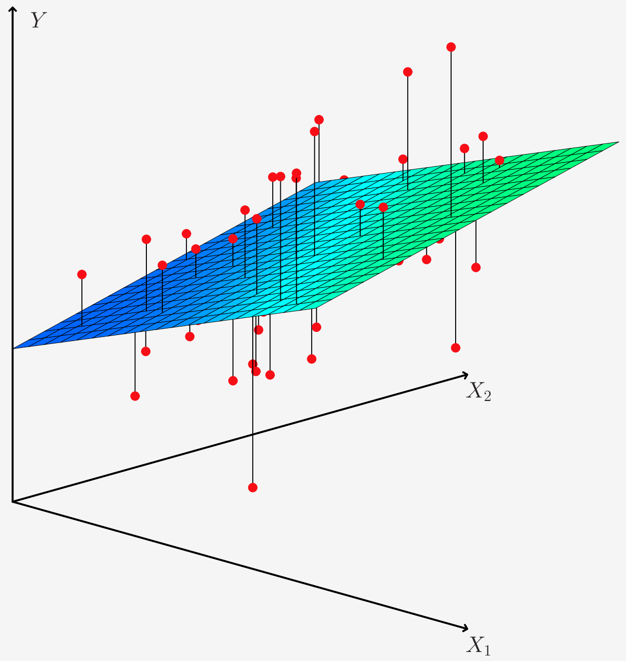
House Sale Price Example:

$Y = 2 + 3X_1 + .5X_2 + 1 \cdot X_3$

SQ FT — 2000

#BEDROOMS — 4

LOT SIZE — 2 ACRES



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

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

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

$$SSE = \sum_{i=1}^{n} \left(y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_p x_{ip} \right) \right)^2$$

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

OK. Back in your groups. 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 produce sales, but our MLR model did not?

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. In reality, higher temperatures cause more people to head to the beach, which in turn increases the chance of shark attacks. Additionally, higher temperatures also cause more people to buy ice cream.

If we ran a multiple linear regression 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

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

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

Covariance and Correlation of Features

One way to discover this relationship between features is to do a correlation analysis. We want to know, if the value of one feature goes up is it likely that the other feature will go up as well? Similarly, we might find that if one feature goes up is it likely that the other feature will go down?

Def: Let X and Y be random variables. The covariance between X and Y is given by

$$\mathrm{Cov}(X,Y) = E[(X - E[X])(Y - E[Y])]$$

Def: The correlation coefficient $\rho(X,Y)$ is a measure between -1 and 1, given by

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

Estimating Covariance and Correlation

We can estimate these relationships from the data using formulas analogous to the sample variance

Def: The sample covariance is given by

$$S^2_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \Big/ n-1$$

Def: The sample correlation coefficient is then given by

$$\hat{\rho}(x,y) = \frac{S^2_{xy}}{\sqrt{S^2_x S^2_y}}$$

Advertising Budget Example

Let's compute the pairwise correlation coefficients for the TV, radio, and newspaper spending features in the advertising data.

In [40]:
```
dfAd[["tv", "radio", "news"]].corr()
```

Out[40]:

	tv	radio	news
tv	1.000000	0.054809	0.056648
radio	0.054809	1.000000	0.354104
news	0.056648	0.354104	1.000000

Question: What do you notice?

Looking Forward

Next time we'll look performing inference on MLR parameters. We'll see how to

- Perform HT to determine if any of the features are related to the response
- Perform HT to determine if a particular subset of features is related to the response
- Extend SLR Goodness-of-Fit measures to the MLR setting
- Perform model selection to get the best lean-and-mean MLR model that we can

For the rest of today we'll look how we can use MLR to explain nonlinear relationships between single-feature data and the response.

Get back in your groups and get out your laptops!

Polynomial Regression

For single-feature data, we can fit a polynomial regression model by casting it as a multiple linear regression where the additional features are powers of the original single-feature, x

Using Residual Plots in Polynomial Regr

Recall that the assumed nature of our true model is

OK! Let's Go to Work!

Get in groups, get out laptop, and open the Lecture 22 In-Class Notebook

Let's:

- Do some stuff!