Class 5: Interaction Effects and Overfitting

Table of Contents / Agenda

1 Interaction Effects

$$Sales = \beta_0 + \beta_1 \times Price + \beta_2 \times Display + \beta_3 \times FeatureAd$$

- If a `Display`, Sales increases by β_2
- If a `Feature Ad`, Sales increase by β_3
- What if there is a `Feature Ad` and `Display` simultaneously?

$$Sales = \beta_0 + \beta_1 \times Price + \beta_2 \times Display + \beta_3 \times FeatureAd + \beta_4 \times (Display \times FeatureAd)$$

$$\text{e.g., } Sales = 100 - 3 \times Price + 5 \times Display + 4 \times FeatureAd + 2 \times (Display \times FeatureAd)$$

- If a `Display`, Sales increases by β_2 $(= 5)$
- If a `Feature Ad`, Sales increase by β_3 $(= 4)$
- If both a `Display` and a `Feature Ad`, sales increase by $\beta_2 + \beta_3 + \beta_4 = 11$

1.1 With a continuous variable

$$Sales = \beta_0 + \beta_1 \times Price + \beta_2 \times Display + \beta_3 \times Price \times Display$$

- What does β_3 represent?

2 Overfitting

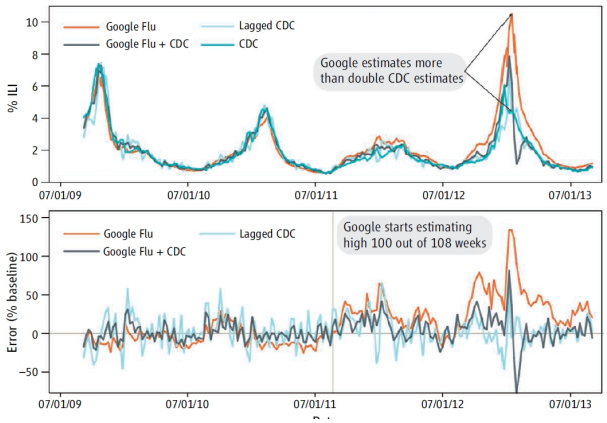
2.1 Google Flu

Google's scientists first announced Google Flu in a Nature article in 2009:

> ... We can accurately estimate the current level of weekly influenza activity in each region of the United States, with a reporting lag of **about one day**.

One report was that Google Flu Trends was able to predict regional outbreaks of flu up to 10 days before they were reported by the CDC

2.1.1 Results



(source: The Parable of Google Flu: Traps in Big Data Analysis)

2.1.2 What Went Wrong?

- Quality of search terms
  - **influenza-like illness**
- Prediction without theory

  $\Rightarrow$ overfitting problem

2.2 Overfitting

2.2.1 Error Term in Regression

- When we think about typical regression model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_K x_{iK} + \varepsilon_i$$

- The error term (ε_i) is supposed to have mean zero
- Not *predictable*
- However, once they are realized, one can often find some pattern in them, which will disappear as more data accumulate
  - e.g., stock prices are supposed to be random-walk; however, from historical data, patterns will pop up

2.2.2 Google Flu and Overfitting

Find the best matches among 50 million search terms to fit 1152 data points

"They ... overfit the data. They had fifty million search terms, and they found some that happened to fit the frequency of the 'flu' over the preceding decade or so, but really they were getting idiosyncratic terms that were peaking in the winter at the time the 'flu' peaks ... but wasn't driven by the fact that people were actually sick with the 'flu',"

(David Lazer, an interview with Science)

2.3 Takeaways

- Be careful of "overfitting", especially when you have a lot of variables
- Conduct out-of-sample validation

2.4 Out-of-Sample Validation

- Use only a subset of data (e.g., 80% of the sample; train dataset) to estimate coefficients
- Then predict $\hat{y}_i$ values for the rest of the sample (validation dataset) using the estimated coefficients and actual data for x_k's as if we do not know actual y_i:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_K x_{Ki}$$

- Compare $\hat{y}_i$ and the actual y_i