IEOR 242 Lab 2

Linear Regression in R

Mengxin Wang

IEOR, UC Berkeley

11/09/2019

Today’s Agenda

(Pair programming, please)

- Linear Regression theory (review)
- General coding notes
- Wine quality Prediction using Linear Regression

Linear Regression

Assumption:

$$\begin{aligned} y &= \beta_0 + \beta_1 * x_1 + \ldots + \beta_n * x_n + \omega \qquad (1) \\ &= \beta^T x + \omega \qquad (2) \end{aligned}$$

White noise: $\omega \sim \mathcal{N}(0, \sigma^2)$
Given samples $\{(x_i, y_i)\}_{i=1}^{n}$, assume X is of full column rank, we can estimate β as

$$\begin{aligned} \hat{\beta}^* &= argmin_\beta \|y - X\beta\|^2 \qquad (3) \\ &= (X^T X)^{-1} X^T y \qquad (4) \end{aligned}$$

Data transformation

In statistics, data transformation is the application of a deterministic mathematical function to each point in a data set — that is, each data point x is replaced with the transformed value $x\prime = f(x)$, where f is a function.

- Linear transformation $f(x) = Ax$
- Logarithm transformation $f(x) = log(x)$
- Exponential transformation $f(x) = exp(x)$

Download

- Lab2.R
- Wine_agg.csv (aggregated across different wineries)
- Wine_disagg.csv

Packages

- dplyr
- ggplot2: for creating graphics
- GGally: extends ggplot2
- car: for VIF

Remember to install all the packages before using them!

```
install.packages(c(”dplyr”, ”ggplot2”, ”GGally”))
```

How to approach coding something

Two types of programming exploration:

- finding which function to use
- understanding how to use a function once you've found it

For 1.:

- Try to target the problem you want to address as succinctly as possible in google.
- Read the description of the function and what it returns

For 2.

- What to do with new functions that you've never seen before?
- Read the arguments carefully, look at examples, try it out and test output

Multicollinearity

Effects of Multicollinearity

1. It will be difficult to find the correct predictors from the set of predictors.
2. It will be difficult to find out precise effect of each predictor.

Side Notes

- Capture non-linear relationships using LR: adding interaction or power terms

Python Resources

- Use numpy and do matrix multiplication from scratch
- Or use packages: sklearn.linear_model, sklearn.metrics