2. (Nonparametric) regression analysis

Regression analysis

egression analysis	2
Main idea	3
Linear regression	4
Nonparametric regression	5
Discrete independent variables	6
Continuous independent variables	7
Bias-variance trade off	8
Local averaging	9
Effects of local averaging	0

Regression analysis

Main idea

- Regression analysis examines the relation between a single dependent variable Y and one or more independent variables X₁,...,X_k.
- Regression analysis describes the conditional distribution of Y given x_1, \ldots, x_k : $f(Y|x_1, \ldots, x_k)$. Usually we describe the mean of this distribution.

See overhead

■ It can be used for:

- describing how Y depends on X_1, \ldots, X_k
- predicting Y from X_1, \ldots, X_k
- inference about the effect of X_1, \ldots, X_k on Y

3 / 10

Linear regression

- Full name: Ordinary least squares multiple linear regression.
- Assumptions of linear regression (see overhead):
 - Data is representative for the population of interest.
 - $E(Y|x_1,\ldots,x_k)$ is a *linear* function of x_1,\ldots,x_k .
 - The variance of $f(Y|x_1, \ldots, x_k)$ does not depend on x_1, \ldots, x_k .
 - $f(Y|x_1,...,x_k)$ is (approximately) normal.

Nonparametric regression

- See section 2.5 of script
- Nonparametric regression does not assume a model (linearity, normality, etc)
- Why consider it?
 - Much weaker assumptions
 - By looking at it we will see its limitations
 - Modern methods of nonparametric regression are emerging

5 / 10

4 / 10

Discrete independent variables

- **E** Recall: Regression analysis describes the conditional distribution $f(Y|x_1, \ldots, x_k)$.
- In very large samples, and if the X's are discrete, we can directly examine this conditional distribution.
- But if there are many independent variables, this becomes problematic:
 - ◆ Three independent variables with 10 possible outcomes already give 10³ = 1000 combinations to look at.
 - \blacklozenge We need a very large data set to have sufficient data at each combination.
 - This is called the "curse of dimensionality".

6 / 10

Continuous independent variables

- If the X's are continuous, we only have one observation for each combination of X's
- Solution:
 - Dissect the range of the X's into a large number of narrow strips
 - Compute average of x and y values in each strip
 - ♦ See R-code

Bias-variance trade off

- To minimize the variance, we want many observations in each strip
- To minimize the bias, we want narrow strips
- We can achieve both if the data set is very large. If this is not the case, or if there are many independent variables, this method is problematic.

8 / 10

7 / 10

Local averaging

- The method with strips is quite rough. We only estimate at a few points
- Solution: Use overlapping strips (moving window):
 - Use each of the *x*-values as midpoint
 - Use either fixed width windows, or windows that contain a fixed number of data points
 - ◆ See R-code

9 / 10

Effects of local averaging

- First few and last few local averages are identical
- Line is rough the average jumps up and down if observations enter and leave the window
- Unusual data values (outliers) have a lot of impact

We can address the 2nd and 3rd problem by weighting:

- Give greater weight to observations close to the center of the window, and small weight to observations close to the edge of the window
- Give small weight to outlying observations

This, and some other techniques, are built into the Lo(w)ess smoother of R. Adding a Loess smoother to a scatterplot is usually helpful for seeing the pattern in the data.

10 / 10