Lecture 16: (More) exploratory data analysis– Non Parametric comparisons and regressions

Prof. Esther Duflo

14.310x

Exploratory Data Analysis: Looking for Patterns before building models

- With RCT, we (often) have a pretty clear hypothesis to test.
- With observational data this may not be the case.
- We want to start getting a sense of what is in our data set
- Early in the semester we discussed how to visualize one distribution
- And started to plot two together: we will start from there!

Combining a continuous distribution and a categorical variable

- Reminder: the basketball players
- We combined the data sets , we can compare pdf, cdf, box plots

Comparing two distributions: Kolmogorov-Smirnov Test

- In analyzing RCT, we have seen how to test the sharp null, and how to test the hypothesis that the treatment has zero effect on average.
- We may also be interested in testing the hypothesis that the distribution of $Y(1)$ and $Y(0)$ are different.
- Kolmogorov-Smirnov statistic. let $X_1, .., X_n$ be a random sample, with CDF F and $Y_1, .., Y_m$ be a random sample, with CDF G
- We are interested in testing the hypothesis

$$H_o : F = G$$

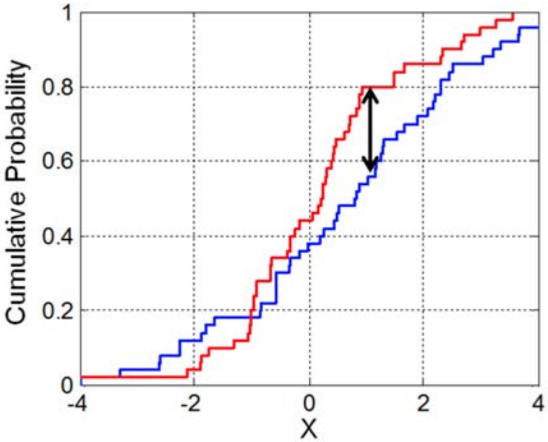
against

$$H_a : F \neq G$$

The statistic

- $D_{nm} = max_x |F_n(x) - G_m(x)|$ where F_n and G_m empirical CDF in the first and the second sample
- Empirical CDF just counts the number of sample point below level x:
$$F_n(x) = P_n(X < x) = \frac{1}{n} \sum_{i=1}^{n} I(X < x)$$

Illustration



First order stochastic dominance: one sided Kolmogorov-Smirnov Test

- We may want to know more, e.g. does the distribution in Treatment first order stochastically dominate the distribution in the control,
- We are interested in testing the hypothesis

$$H_o : F = G$$

against

$$H_a : F > G$$

(which would mean that G FSD F).
- The one sided KS statistics is: $D^+_{nm} = max_x[F_n(x) - G_m(x)]$ (remove the absolute value).

Asymptotic distribution of the KS statistic

Under H_o, the limit of KS as N and N' go to infinity is 0, so we want to compare the KS statistics to 0. So we will reject the hypothesis if the statistics is "large" enough.
The key observation that underlies the KS testing is that, under the null, the distribution of

$$(\frac{nm}{n+m})^{\frac{1}{2}} D_{nm}$$

does not depend on the unknown distribution in the samples: it has a known distribution (KS) , with associated critical values. Therefore we reject the null of equality if $D_{nm} > C(\alpha)(\frac{nm}{n+m})$, where $C(\alpha)$ are critical values which we find in tables (and R knows).
We can test this with the Basketball players, using the ks.test command in R.

Note: you could use the KS test in ONE sample

To test, for example, whether the sample follow some specific distribution (e.g. a normal one).

$$D_n = max_x |F_n(x) - F(x)|$$

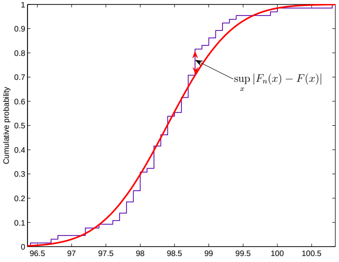


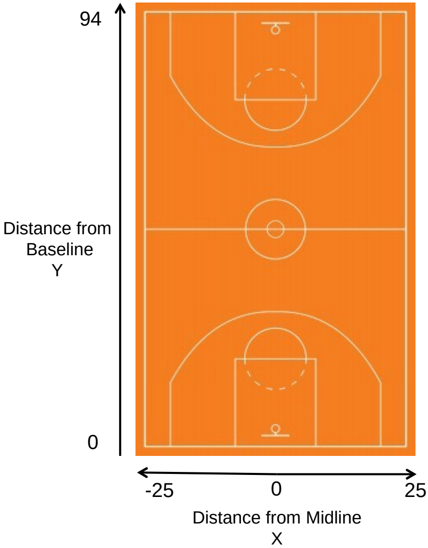
Figure 13.2: Kolmogorov-Smirnov test statistic.

Reject if $\sqrt(n) D_n > K(\alpha)$
We can do this in R with ks.test, again we can test this with Steve Curry.

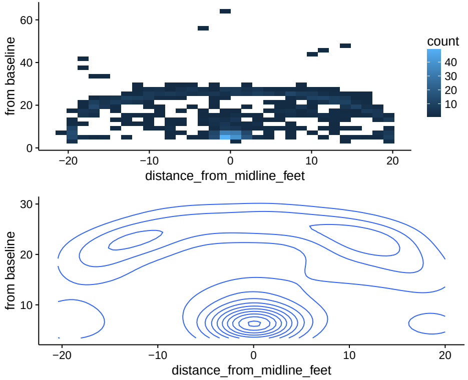
Representing joint distributions

- Suppose we want to represent the distribution of successful attempts by location
- There are actually two distances to consider: distance from baseline, and distance from the sideline
- If we plot each of them separately, what do we get?

A basketball court



A histogram of the joint density–or the map of a basketball court?



Now we see pretty clearly that there is bunching at the 3pt line!

Two continuous variables

- Refer to the R code NP.R for a way to approach two variables, using the relationship between earnings and wages.
- Now we need to go under the hood– How does R estimate the non-parametric function between two variables. We will start with something ggplot does not do, but could.... Kernel regression.

Non Parametric (bi-variate) Regression

You have two random variable, X and Y and express the conditional expectation of X given Y as : $E[Y|X] = g(X)$
Therefore, for any x, and y,

$$y = g(x) + \epsilon$$

where ϵ is the prediction error.
You may think that this relationship is causal or not. Problem is to estimate $g(x)$ without imposing a functional form.

The Kernel regression: A common non-parametric regression

g(x) is the conditional expectation of y given x.

$$E(Y|X = x) = \int yf(y|x)dy$$

By Bayes's rule:

$$\int yf(y/x)dy = \int \frac{yf(x, y)dy}{f(x)} = \frac{\int yf(x, y)dy}{f(x)}$$

Kernel Estimator

Kernel estimator replace $f(x,y)$ and $f(x)$ by their empirical estimates.

$$\hat{g}(x) = \frac{\int y\hat{f}(x,y)dy}{\hat{f}(x)}$$

- Denominator: estimating the density of x (we have seen this!)

  $$f(\hat{x}) = \frac{1}{N * h}\sum_{i=1}^{n} K(\frac{x - x_i}{h}),$$

  where h is a positive number (the bandwith) is the kernel estimate of the density of x. $K(.)$ is a density, i.e. a positive function that integrates to 1
  It is a weighted proportion of observations that are within a distance h of the point x.

Kernel Estimator

Kernel estimator replace $f(x, y)$ and $f(x)$ by their empirical estimates.

$$\hat{g}(x) = \frac{\int y\hat{f}(x, y)dy}{\hat{f}(x)}$$

- Numerator

$$\frac{1}{N * h} \sum_{i=1}^{n} y_i K(\frac{x - x_i}{h})$$

Combine the two

$$\hat{g}(x) = \frac{\sum_{i=1}^{n} y_i K(\frac{x-x_i}{h})}{\sum_{i=1}^{n} K(\frac{x-x_i}{h})} \tag{1}$$

$\hat{g}(x)$ is a weighted average of Y over a range of points close to x.
The weights are declining for points further away from x.
In practice, you choose a grid of points (ex. 50 points) and you calculate the formula given in equation 1 for each of these points.

Large sample properties

- as h goes to zero, bias goes to zero
- as nh goes to infinity, variance goes to zero.
- So as you increase the number of observation, you "promise" to decrease the bandwidth

Choices to make

- Choice of Kernel
  1. Histogram: $K(u) = 1/2$ if $\mid u \mid \leq 1$, $K(u) = 0$ otherwise.
  2. Epanechnikov $K(u) = \frac{3}{4}(1 - u^2)$ if $\mid u \mid \leq 1$ $K(u) = 0$ otherwise
  3. Quartic
     $K(u) = (\frac{3}{4}(1 - u^2))^2$ if $(u \leq 1)$, $K(u) = 0$ otherwise
- Choice of bandwidth : Trade off Bias, and Variance
  - A large bandwidth will lead to more bias (as we are missing important features of the conditional expectation).
  - A small bandwidth will lead to more variance (as we start to pick up lots of irrelevant ups and downs)

Cross Validation

One way to formalize this choice is cross validation.
First, define for each observation i define the prediction error as:

$$e_i = y_i - \hat{g(x_i)}$$

and the leave out prediction error as:

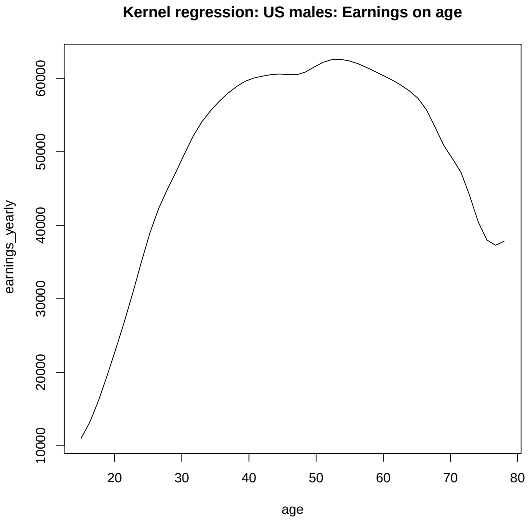
$$e_{i,-i} = y_i - \hat{g_{-i}}(x_i)$$

where $\hat{g_{-i}}(x_i)$ is the prediction of y based on kernel regression using all the observations except i.
An optimal bandwidth will minimize

$$CV = \frac{1}{N} \sum_{i=1}^{N} e_{i,-i}^2$$

(or often in practice $CV = \frac{1}{N} \sum_{i=1}^{N} e_{i,-i}^2 M(X)$) where $M(X)$ is a trimming function to avoid influence of boundary points)

Kernel regression with optimal bandwidth



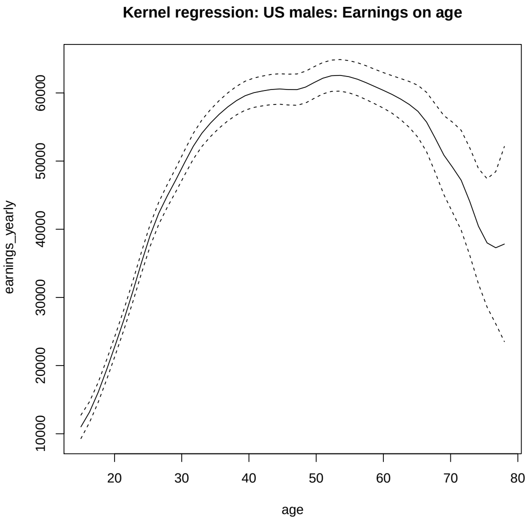
Confidence bands

$y_i = g(X_i) + e_i$ and $E[e_i|X_i] = 0$
$e_i^2 = \sigma_i^2(X_i) + \eta_i$ and $E[\eta_i|X_i] = 0$
So a Kernel estimate of $\sigma_i^2(X_i)$ is :

$$\hat{\sigma}^2(x) = \frac{\sum_{i=1}^{n} e_i^2 K(\frac{x-x_i}{h})}{\sum_{i=1}^{n} K(\frac{x-x_i}{h})} \tag{2}$$

Point-wise confidence interval can be drawn using this estimate.

Kernel regression with confidence bands



Other non parametric methods

- Series estimation (approximate the curves by polynomes); splines (polynomes with knots)
- Local linear regression(instead of taking the mean, in each interval, take predicted value from a regression (Loess).