MA 575 Linear Models:

Cedric E. Ginestet, Boston University

Scatterplots and LOESS

Week 2, Lecture 1

1 Revision: Conditional Distributions

1.1 Conditional Probability

A classical example of conditional probability can be obtained by considering **discrete** random variables. In this case, we are concerned with **probability mass functions**, as opposed to the probability density functions of *continuous* random variables.

Let two coins, C_1 and C_2, such that C_1 is a *fair* coin and C_2 is a *double-sided* coin with heads on both sides. The random variable of interest, denoted $X \in \{H, T\}$, takes either a value of heads or tails. Observe that X can only take a single value. Thus, the value of X depends on the choice of coin. This **dependency** is formally defined using conditional probability, by treating the choice of coins as another random variable, $Y \in \{C_1, C_2\}$. Thus, the **conditional probability** of obtaining *heads* or *tails*, given that we are using C_1 are identical

$$\mathbb{P}[X = H|Y = C_1] = \mathbb{P}[X = T|Y = C_1] = \frac{1}{2}.$$

By contrast, if we are using the double-sided coin C_2, we have

$$\mathbb{P}[X = H|Y = C_2] = 1, \qquad \text{and} \qquad \mathbb{P}[X = T|Y = C_2] = 0.$$

Now, what is the **marginal probability** of obtaining heads? That is, once we have taken into account the fact that we may choose over two different coins, what is the overall probability of obtaining heads? This only makes *if we can assign a probability mass function to the choice of coins*. Let us assume that the chances of choosing C_1 or C_2 are identical, such that

$$\mathbb{P}[Y = C_1] = \mathbb{P}[Y = C_2] = \frac{1}{2}.$$

Then, we can use the **law of total probability**, in order to compute the marginal probability of obtaining heads,

$$\mathbb{P}[X = H] = \sum_{j=1}^{2} \mathbb{P}[X = H|Y = C_j]\mathbb{P}[Y = C_j] = \frac{1}{4} + \frac{2}{4} = \frac{3}{4}.$$

This can be easily verified using a *decision tree*.

1.2 Conditional Probability Density Functions

Every pdf is defined **conditionally** on the values of some parameters. That is, when we say that a variable Y is normally distributed, as in

$$Y \sim N(\mu, \sigma^2),$$

we are stating a dependence of Y on the parameters μ and σ^2. It suffices to assign particular values to these parameters in order to produce specific probabilities for any given values of Y, such that when Y is evaluated at its mean, we have

$$\mathbb{P}[Y = 0|\mu = 0, \sigma^2 = 1] = \frac{1}{\sqrt{2\pi}}.$$

Naturally, these parameters could also be treated as random variables, by assigning them probability distributions. This is exactly what we do in a **Bayesian framework**. The mean parameter, μ, for instance, could also be given a normal distribution, such that

$$\mu \sim N(\xi, \tau^2).$$

Similarly, in the case of linear regression, which is central to this course, we have assumed that the predictor X (or **independent variable**) is fixed, or **known**. However, for convenience, we are still using the notation, inherited from the '*possibility*' of considering this variable as random. Thus, when we state that the expectation of Y is dependent on X, as in

$$\mathbb{E}[Y|X = x] = \beta_0 + \beta_1 x,$$

we assume that X has been **measured without noise**. This is especially the case, when the variable X is the one **manipulated by the experimenter**. In an experimental setting, we are:

i. *Experimentally fixing* the values of X,

ii. *Measuring* the values of Y.

The assumption that X is measured without noise is crucial in the context of simple linear regression. Other models may make different assumptions, and model the randomness of X, if the values of that variable have been potentially corrupted with noise.

2 Scatterplots

2.1 LOESS: Non-parametric Regression

LOWESS means LOcally Weighted Scatterplot Smoothing. A later development was LOESS, which stands for LOcal regrESSion. This is an algorithm, which proceeds as follows:

i. Select a **bandwidth**, **smoothing parameter**, or **span**, denoted f, and comprised between 0 and 1. When $f = 0$, we obtain a very rough LOESS, which tries to wiggle to visit every point. This is a case of **under-smoothing**. When $f = 1$, we obtain a very smooth line, which is straight. This is a case of **over-smoothing**.

ii. Given a set of points, $\widetilde{x}_1, \ldots, \widetilde{x}_m$, which "*covers* the range" of X, we identify the nf points, which are the closest to each $\widetilde{x}_j$. This is a k-nearest neighbor method.

iii. We next can fit a weighted least-squares (WLS) model for each of the nf nearest neighbors of $\widetilde{x}_j$, where the weights are determined by the distance from $\widetilde{x}_j$.

iv. Repeat this for every $\widetilde{x}_j$, in the cover of the range of X.

2.2 Non-parametric Estimation of Variance Function

A similar method can be used to estimate the variance function of a pair of variables, using a **locally approximated normal kernel**.

2.3 Box-plots and Quartiles

When the function $F : \mathbb{R} \mapsto [0, 1]$ is (i) *continuous* and (ii) *strictly monotonic*, we can define

$$Q(p) := F^{-1}(p),$$

for some real number $p \in [0, 1]$. That is, $\forall\ p \in [0, 1]$, there exists a unique $x \in \mathbb{R}$, such that $F(x) = p$, and we define that x as $Q(p) = F^{-1}(p)$.

In general, however, CDFs need not be continuous and strictly monotonic. (Note that CDFs are always monotonic, but not necessarily **strictly** monotonic.) However, it is possible to uniquely define the quantile function of any random variable, in the following manner,

$$Q(p) := \inf_{x:F(x)\geq p} x,$$

for every $p \in [0, 1]$. Then, the **quartiles** of a random variables are the quantiles corresponding to the values

$$p := \left\{\frac{1}{4}, \frac{2}{4}, \frac{3}{4}, \frac{4}{4}\right\},$$

where the second quartile is simply the **median**, which may also be expressed as the 50^{th} **percentile**.