Notes on Law of Iterated Expectations

Cyrus Samii

The law of iterated expectations is the foundation of many derivations and theorems in applied statistics. I provide the basic statement of the law and then illustrate how it applies to some important results.

A basic statement is as follows:

$$\mathrm{E}\left(Y\right) = \mathrm{E}\left(\mathrm{E}\left(Y|X\right)\right).$$

Here is a proof for the continuous case:

$$\begin{aligned}\mathrm{E}\left(\mathrm{E}\left(Y|X\right)\right) &= \int \mathrm{E}\left(Y|X\right) f_X(x)dx = \int \left(\int y f_{Y|X}(y|x)dy \right) f_X(x)dx \\ &= \int\int y f_{X,Y}(x,y)dxdy = \int y \left(\int f_{X,Y}(x,y)dx \right) dy \int y f_Y(y)dy = \mathrm{E}\left(Y\right),\end{aligned}$$

and for the discrete case,

$$\begin{aligned}\mathrm{E}\left(\mathrm{E}\left(Y|X\right)\right) &= \sum_x \mathrm{E}\left(Y|X=x\right)p(x) = \sum_x \left(\sum_y y p(y|x) \right) p(x) \\ &= \sum_x \sum_y y p(x,y) = \sum_y y \sum_x p(x,y) = \sum_y y p(y) = \mathrm{E}\left(Y\right).\end{aligned}$$

Example 1. Suppose $Y|X \sim N(2X, 1)$ and $X \sim$ Multinomial$(.25, .25, .5)$. This defines a joint distribution for Y and X (since $f(Y|X)f(X) = f(X,Y)$). Take a large number of draws from this joint distribution. Average the Y realizations. Then, bin the Y realizations by their associated X values, take the average of the Y realizations within each bin, and finally take the probability-weighted average over the bins. This should equal the mean of the Ys. Here is a demonstration with a finite (though rather large) sample using R:

```
> n <- 5000
> p <- c(.25,.25,.5)
> x <- t(rmultinom(n,1,p))%*%c(1,2,3)
> y <- rnorm(n, mean=2*x, sd=1)
> mean(y)
[1] 4.500761
> tapply(y, x, mean)
       1        2        3
1.994880 3.973878 6.033394
> tapply(y, x, mean)%*%as.matrix(p)
         [,1]
[1,] 4.508887
```

As the example illustrates, $\mathrm{E}\left(Y|X\right)$ is itself a random variable. In the example, we have,

$$\mathrm{E}\left(Y|X\right) = \begin{cases} 2 & \text{if } X = 1 \\ 4 & \text{if } X = 2 \\ 6 & \text{if } X = 3 \end{cases},$$

where the probability of any of those $\mathrm{E}(Y|X)$ realizations is equal to the probability of realizing $X = 1$ vs $X = 2$ vs $X = 3$, respectively. Thus, $\mathrm{E}(Y|X)$ is random with respect to the variation in X, and so the outer expectation is taken with respect to this source of random variation—namely, the density of X.

Iterated expectations can be taken with nested conditioning sets. For example, we have

$$\mathrm{E}(Y|X) = \mathrm{E}(\mathrm{E}(Y|X,Z)|X).$$

The proof is no more complicated than what we have above—just substitute in the appropriate conditional densities (e.g., replace f_Y with $f_{Y|X}$). What is important is the statistical interpretation, which typically looks at the conditioning sets in terms of their information content. In these terms, the conditioning sets are nested: $X \subseteq (X,Z)$. In a sense, the variability in $\mathrm{E}(Y|X,Z)$ is more "fine grained" than the variability in $\mathrm{E}(Y|X)$ given the additional "resolution" contributed by the information on Z. The outer expectation retains the variability with respect to X but marginalizes over Z. That is, $\mathrm{E}(Y|X)$ is "taking an average" over the fine grain that we would observe if we *could* condition on (X,Z) but that we *cannot* do because we only have X. Hence the equality.

Example 2 (Martingales): Suppose a random walk, $Y_t = Y_{t-1} + u_t$ where $u_t \sim iid(0, \sigma_t^2)$. Suppose we have an information set, $I_t = \{u_t, u_{t-1}, ...\}$. Then, with I_{t-1} the value of Y_{t-1} is fully determined, and we are able to form an expectation about Y_t as follows: $\mathrm{E}(Y_t|I_{t-1}) = \mathrm{E}(Y_{t-1} + u_t|I_{t-1}) = Y_{t-1}$. $\{Y_t, I_t\}$ is an example of a martingale process. Now, what would be $\mathrm{E}(Y_t|I_{t-2})$, our expectation of the price at time t given information only up to $t-2$? Since $I_{t-2} \subseteq I_{t-1}$, by application of LIE, $\mathrm{E}(Y_t|I_{t-2}) = \mathrm{E}(\mathrm{E}(Y_t|I_{t-1})|I_{t-2}) = \mathrm{E}(Y_{t-1}|I_{t-2}) = Y_{t-2}$. As before, the outer expectation is "averaging" over the finer grain I_{t-1} information to generate an expectation given our coarser I_{t-2} information. By induction, we see that $\mathrm{E}(Y_t|I_{t-k}) = Y_k$, which is why a martingale such as this is sometimes called an "honest" process.

Example 3 (Propensity score theorem): The propensity score theorem is an important result for methods used to correct for problems of unequal sample selection or treatment assignment.[1] Each unit i is assigned a binary treatment, $D_i \in \{0,1\}$. Suppose each unit i has potential outcomes Y_{0i} and Y_{1i} which are observed when $D_i = 0$ and $D_i = 1$, respectively. Suppose as well that each unit can be characterized by covariates, X_i, such that "conditional independence" holds, $\{Y_{0i}, Y_{1i}\} \perp\!\!\!\perp D_i|X_i$. That is, for units with common X_i, treatment assignment is random (or "unconfounded") with respect to potential outcomes. Finally, let $p(X_i) = \Pr(D_i = 1|X_i)$. The propensity score theorem states that under these conditions, $\{Y_{0i}, Y_{1i}\} \perp\!\!\!\perp D_i|p(X_i)$. The proof is based on a straightforward application of LIE with nested conditioning:

$$\begin{aligned}
\Pr(D_i = 1|\{Y_{0i}, Y_{1i}\}, p(X_i)) &= \mathrm{E}(D_i|\{Y_{0i}, Y_{1i}\}, p(X_i)) \\
&= \mathrm{E}(\mathrm{E}(D_i|\{Y_{0i}, Y_{1i}\}, p(X_i), X_i)|\{Y_{0i}, Y_{1i}\}, p(X_i)) && \text{(by LIE)} \\
&= \mathrm{E}(\mathrm{E}(D_i|\{Y_{0i}, Y_{1i}\}, X_i)|\{Y_{0i}, Y_{1i}\}, p(X_i)) \\
&= \mathrm{E}(\mathrm{E}(D_i|X_i)|\{Y_{0i}, Y_{1i}\}, p(X_i)) && \text{(by conditional independence)} \\
&= \mathrm{E}(p(X_i)|\{Y_{0i}, Y_{1i}\}, p(X_i)) = p(X_i),
\end{aligned}$$

Thus, under conditional independence, $\Pr(D_i = 1|\{Y_{0i}, Y_{1i}\}, p(X_i)) = p(X_i) = \Pr(D_i = 1|p(X_i))$. Since $\Pr(D_i = 1|p(X_i))$ completely characterizes the distribution of D_i conditional on $p(X_i)$, we can conclude that $D_i \perp\!\!\!\perp \{Y_{0i}, Y_{1i}\}|p(X_i)$. The usefulness of the propensity score theorem is that it reduces the dimensionality of the problem of correcting for non-random selection or treatment assignment. If X_i is of high dimension, then conditioning on X_i directly—e.g. by creating stratification cells—could be hard due to sparse data over values of X_i.

[1] P Rosenbaum and DB Rubin, 1983, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, 70(1):41-55. The proof here follows the presentation in JD Angrist and JS Pischke, 2008, *Mostly Harmless Econometrics*, Princeton: Princeton University Press, pp. 80-81.

The propensity score theorem states that the information from X_i can be aggregated into a single dimension score, $p(X_i)$, and that conditioning on this is sufficient to correct for non-random selection or treatment assignment. The conditioning can take the form of stratification on discretized values of the propensity score or specifying a model that includes the propensity score as a covariate.

Example 4 (Inverse probability weighted mean model): Let X_i be a set of covariates characterizing units i, and let $D_i \in \{0, 1\}$ be an indicator for whether i is observed. Suppose we are interested in computing $\mathrm{E}(Y_i)$ for the population of units indexed by i, but that in fact each unit i is observed with probability $p(X_i) = \mathrm{E}(D_i|X_i)$. In other words, we want to use a non-representative sample of a population (namely, $D_1Y_1, D_2Y_2, ...$) to compute the mean of the population ($Y_1, Y_2, ...$). Finally, suppose that a form of conditional independence holds such that $Y_i \perp\!\!\!\perp D_i|X_i$. We want to show that the inverse probability weighted (IPW) mean of the observed data, $\mathrm{E}(D_i\frac{Y_i}{p(X_i)})$, is unbiased for the population mean, $\mathrm{E}(Y_i)$.

$$
\begin{aligned}
\mathrm{E}(D_i\frac{Y_i}{p(X_i)}) &= \mathrm{E}(\mathrm{E}(D_i\frac{Y_i}{p(X_i)}|Y_i, X_i)) && \text{(by LIE)} \\
&= \mathrm{E}(\frac{Y_i}{p(X_i)}\mathrm{E}(D_i|Y_i, X_i)) \\
&= \mathrm{E}(\frac{Y_i}{p(X_i)}\mathrm{E}(D_i|X_i)) && \text{(by conditional independence)} \\
&= \mathrm{E}(\frac{Y_i}{p(X_i)}p(X_i)) = \mathrm{E}(Y_i).
\end{aligned}
$$

Hopefully you should notice the similarities to the previous example. The conditional independence assumption that we used here is essentially the same as the one used in the previous example. In this way, we see that that conditional independence provides multiple methods through which propensity scores can be used to correct for non-random selection or treatment assignment. This observation is fundamental in the development of "doubly robust" estimators.[2] The unbiasedness of the IPW mean is the basis of survey weighting, although survey theorists typically use a finite population characterization to demonstrate this.[3]

LIE is also used for other important results, including the demonstration that all random variables can be decomposed into the sum of a conditional mean and an uncorrelated error term with mean zero, and that the variance of all random variables can be decomposed into "explained variance" and "residual variance." These results are the basis for using linear regression to estimate basic relations between variables even if the relationship is not inherently linear. See Angrist and Pischke (2008), pp. 32-34.[4]

[2]H Bang and JM Robins, 2005, "Doubly robust estimation in missing data and causal inference models," *Biometrics*, 61:962-972.

[3]DG Horvitz and DJ Thompson, 1952, "A generalization of sampling without replacement from a finite universe," *Journal of the American Statistical Association*, 47:663-685.

[4]See fn. 1.