SST is the total amount of variation of Y around its mean. The distance from each point, to the average line for Y. SSE is the amount of variation around the regression line. So basically the variation that is not expained by the line. The distance from each point to the line. So the line is an attempt to show that Y varies in some sort of trend in relation to X, because it's variation around the line is smaller than its overall variation. So the proportion of Y's variance around the line (the variation in X) to its variance around its mean, is 1 - R^2. So the smaller the variation around the line, the higher the R^2, because R^2 is a measure of how much of the overall variation in Y is explained by the variation in X. So as X changes (goes from left to right) Y does as well.

So as you add more and more explanatory variables, the R^2 will rise automatically because there is just more noise which in some cases could accidentally "explain" some of the leftover variation in Y.

Bayesian data analysis has two foundational ideas. The first idea is that Bayesian inference is reallocation of credibility across possibilities. The second foundational idea is that the possibilities, over which we allocate credibility, are parameter values in meaningful mathematical models.

Marcov Structure means only the most recent state matters.

Bayesian:

Bayesian data analysis begins with a descriptive model, just as in classical statistics. The descriptive model has meaningful parameters that describe trends in the data. Unlike classical methods, Bayesian analysis yields a complete distribution over the joint parameter space, revealing the relative credibility of all possible combinations of parameter values. Decisions about parameter values of special interest, such as zero, can be made directly from the derived distribution.

When the rule is applied to parameters and data, it can be written as follows: Posterior = Likelihood * Prior / Evidence, or, P(y|D) = P(D|y) * P(y) / P(D). Where D is the observed data and y is a vector of parameters in the descriptive model. The posterior distribution, p(y|D), specifies the relative credibility of every combination of parameters given the data. Because the range of parameter values defines the complete space of possible descriptions, the distribution of credibility sums to 1 and is tantamount to a probability distribution. The posterior distribution provides the most complete information that is mathematically possible about the parameter values given the data (unlike the point estimate and confidence interval in classical statistics, which provide no distributional information).

Specifically, what researchers want to know is the parameter values that are credible, given the observed data. In particular, researchers may want to know the viability of a null hypothesis (e.g., zero correlation between two variables or zero difference in mean scores across groups) given the data, p(Ho|D). However, traditional methods based on NHST, also labeled frequentist statistics to distinguish them from Bayesian statistics, tell us the probability of obtaining the data

in hand, or more extreme unobserved data, if the null hypothesis were true, p(D|Ho) (Aguinis, Werner, et al., 2010). Unfortunately, p(Ho|D) 6¼ p(D|Ho). As noted by Cohen (1994), a test of statistical significance ‘‘does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does!’’ (p. 997). In short, what we want to know is the credibility of candidate parameter values given the data that we actually observed.

Bayesian analysis determines what can be inferred about parameter values given the actually observed data.

The prior distribution is a continuous mathematical function indicated by the black curves, but it is illustrated with superimposed histograms because the parameter distribution will be represented by using a very large (e.g., 250,000) representative random sample from the parameter space. Thus, the Bayesian analysis will reallocate the very large set of representative parameter values from the prior distribution to a posterior distribution, illustrated by histograms of the representative values. For any fixed set of data and prior distribution, there is one true posterior distribution, represented by a very large representative sample of parameter values. The larger the representative sample, the higher resolution picture we have of the true posterior.

Computing the value of p(D) can be difficult because it is actually an integral over the parameter space: $p(D) = \int p(D|y)p(y)dy$. For many years, Bayesian analysis was confined to mathematical forms that could be analytically integrated or analytically approximated. Fortunately, new computer-based methods have emerged that allow Bayesian analysis to be computed flexibly and for very complex models. The new method is called Markov chain Monte Carlo (MCMC).

The new method is called Markov chain Monte Carlo (MCMC). The idea is to accurately approximate the posterior distribution by a very large representative random sample of parameter values drawn from the posterior distribution. From this very large sample of representative parameter values, we can determine the mean or modal parameter value, the quantiles of the parameter distribution, its detailed shape, and the forms of trade-offs between values of different parameters. What makes this approach possible is that MCMC methods generate the random sample without needing to compute the difficult integral for p(D). Moreover, recent advances in algorithms and software have made it possible for a researcher merely to specify the form of the likelihood function and prior distribution, and the software is able to apply MCMC methods

SO the likelihood part comes from the multilinear estimation using the parameters of mean and variance. The prior is there too. Now the denominator is by Marcov chains?

Chapter 18:

http://www.indiana.edu/kruschke/BMLR/