Using Bayes' Theorem to Develop CERs – Extending the Gaussian Model

Christian Smart, Ph.D.

David Jo

Galorath Federal

Abstract

Bayes' Theorem is a mathematical method for combining prior experience with new data. It is extremely important in leveraging limited information, which is often the case in cost estimating. This paper is an extension of a previous ICEAA paper that dealt with the application of Bayes' Theorem to cost estimating. In this update we show how the Bayesian approach for linear models can be extended for different and more realistic assumptions.

Introduction

Historically, parametric methods in cost estimating have used frequentist techniques. Frequentist statistics is the classical method that uses a sample of data as inputs. If you have taken Statistics 101 in college, most if not all the class was oriented towards this approach. For example, traditional linear and nonlinear regression analysis is a frequentist approach. The challenge with this method is that it requires a large amount of data. Statisticians have conducted numerous studies using random data and have concluded that you need 50 data points for a regression analysis with 10 additional data points for every independent variable you want to include. For example, if you want to include three independent variables in your analysis, you need 80 data points. The number of highly specialized systems used in the Department of Defense and NASA means that we typically have nowhere near that much. For example, the Missile Defense Agency has only developed a handful of different kill vehicles, and NASA has only developed a few crewed launch vehicles. When looking at truly applicable data, the sample size shrinks even further – when considering launch vehicles, the primary systems that NASA has completed have been those for the Apollo and Shuttle programs. The Apollo program began in the 1960s, and the Shuttle program began in the 1970s. Thus, there are no directly applicable historical data points within the last 40 years. Considering the changes that have taken place in the realm of technology since then, there really is no applicable historical data at all for these systems.

For small data sets like these, Bayesian methods can help provide more accurate estimates. Bayesian methods leverage all your experience, making them less subject to being overwhelmed by noise. This prior experience can be subjective or objective. The objective data could involve the use of similar data that is not directly applicable.

This approach has proven to be successful in a multitude of applications. Bayesian techniques were used in World War II to help crack the Enigma code used by the Germans, thus helping to shorten the war. John Nash's equilibrium for games with incomplete or imperfect information is a form of Bayesian analysis (John Nash's life was

portrayed in the film A Beautiful Mind). Actuaries have used Bayesian methods for over 100 years to set property and casualty insurance premiums. Bayesian voice recognition researchers applied their skills as leaders of the portfolio and technical trading team for the Medallion Fund, a $5 billion hedge fund which has averaged annual returns of 35% after fees since 1989.

As you can see, the Bayesian method has been used in several applications in which there is money on the line. I place a high degree of confidence in methods that have proven themselves under such circumstances. They are not mere "ivory tower" exercises without practical application. The practitioners who have used these methods have risked significant sums of money, literally billions of dollars, in using these methods. Thus, many of the users of Bayes' Theorem have "skin in the game" (Taleb 2018).

A previous paper on this subject by one of the authors of this paper dealt with the normal/Gaussian model (Smart 2014). This paper focused on the context of log-transformed ordinary least squares regression and assumed that both the log of the prior information and that the log of the new information is normally distributed. In addition, we assumed that the variance is a known, fixed quantity. The prior experience is typically based on a large amount of information, either objective or subjective, and thus the assumption of normality for the prior information is not suspect. However, we typically do not know the variance – indeed we estimate it with the square of the standard error of the prior model in the case of objective prior data. Even worse, the "known" variance is treated by the modeling process as the square of the standard error of the estimate of the regression equation based on a small sample, which is not an accurate estimate of the true variance. Also, the reason why we are using Bayes' Theorem is that we have a small amount of data. In this case, when we don't know the variance and we have a small amount of data, the new information should be assumed to follow a Student's t-distribution.

In the previous paper on Bayesian regression, we showed that the resulting estimate of a Bayesian approach to regression is a weighted average of the regression coefficients for the two data sets. The weights are determined by the uncertainty about the coefficients. The impact of using the normal model with known variance is that we will typically assign too much weight to the parameters of the smaller data set – the small data set will provide a better fit than the true underlying population, so the standard error of the estimate will be significantly lower than the overall population.

In this paper we extend the application of Bayes' Theorem for the normal linear model to the case of unknown variance, and to the case of a Student's t-distribution for new information. We make use of Markov Chain Monte Carlo simulation as the ability to analytically model these assumptions breaks down as we generalize our assumptions. We also show how the models under these various assumptions can be used for prediction and how to incorporate uncertainty in the estimates derived from these methods.

Bayes' Theorem

In this section we provide a review of Bayes' theorem. The distribution of the model given values for the parameters is called the *model distribution*. *Prior* probabilities are assigned to the model parameters. After observing data, a new distribution, called the *posterior* distribution, is developed for the parameters, using Bayes' Theorem.

The conditional probability of event A given event B is denoted by Pr ($AA|BB$).

In its discrete form, Bayes' Theorem states that

$$PPPP(AA|BB) = \frac{PPPP(AA)PPPP(BB|AA)}{PPPP(BB)}$$

For example, consider testing for illegal drug use. Many people have had to take such a test as a condition of employment with the federal government or with a government contractor. What is the probability that someone who fails a drug test is not a user of illegal drugs? Bayes' theorem can be used to answer such questions.

Suppose that 95% of the population does not use illegal drugs. Also suppose that the drug test is highly accurate. If someone is a drug user, it returns a positive result 99% of the time. If someone is not a drug user, the test returns a false positive only 2% of the time.

In this case: A is the event that someone is not a user of illegal drugs, and B is the event that someone test positive for illegal drugs. The complement of A, denoted A', is the event that someone is a user of illegal drugs.

From the law of total probability,

$$PPPP(BB) = PPPP(BB|AA)\ PPPP(AA) + PPPP(BB|AA')\ PPPP(AA')$$

Thus Bayes' Theorem in this case is equivalent to:

$$PPPP(AA|BB) = \frac{PPPP(BB|AA)\ PPPP(AA)}{PPPP(BB|AA)\ PPPP(AA) + PPPP(BB|AA')\ PPPP(AA')}$$

Plugging in the appropriate values

$$PPPP(AA|BB) = \frac{00.0000(00.9999)}{00.0000(00.9999) + 00.9999(00.0099)} \approx 0022.22\%$$

Thus, even with accurate drug tests it is easy to obtain false positives. This is a case of inverse probability, a kind of statistical detective work where we try to determine whether someone is innocent or guilty based on revealed evidence. More typical of the kind of problem that we want to solve is the following: We have some prior evidence or opinion about a subject, and we also have some direct empirical evidence. How do we take our prior evidence, and combine it with the current evidence to form an accurate estimate of a future event?

It's simply a matter of interpreting Baye's Rule. $PPPP(AA)$ is the probability that we assign to an event before seeing the data. This is called the *prior* probability. $PPPP(AA|BB)$ is the probability after we see the data. This is called the *posterior* probability. $PPPP(BB|AA)/PPPP\ (BB)$ is the probability of the seeing these data given the hypothesis. This is the *likelihood*.

Bayes' Rule can be re-stated as

$$Posterior \propto Prior*Likelihood$$

An example of this application of Bayes' Theorem can be found in the Monty Hall Problem. This is based on the television show *Let's Make a Deal*, whose original host was Monty Hall. In this version of the problem, there are three doors. Behind one door is a car. Behind each of the other two doors is a goat. You pick a door. Monty, who knows what is behind the doors, then opens one of the other doors that has a goat behind it. Suppose you pick door #1. He then opens door #3, showing you the goat behind it, and ask you if you want to pick door #2 instead. See Figure 1. Is it to your advantage to switch your choice?

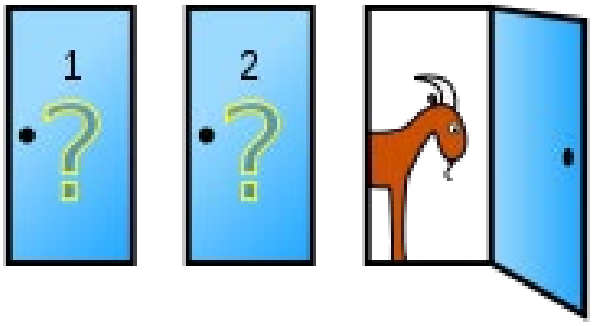

Figure 1. The Monty Hall Problem.

To solve this problem, let A_1 denote the event that the car is behind door #1, A_2 the event that the car is behind door #2, and A_3 the event that the car is behind door #3. Your

original hypothesis is that there was an equally likely chance that the car was behind any one of the three doors. Thus the prior probability, before the third door is opened, that the car was behind door #1, which we denote $\boldsymbol{PPPP}(\boldsymbol{AA}_{11})$, is 1/3. Also, $\boldsymbol{PPPP}(\boldsymbol{AA}_{00})$ and $\boldsymbol{PPPP}(\boldsymbol{AA}_{33})$ are also equal to 1/3.

Once you picked door #1, you were given additional information. You were shown that a goat is behind door #3. Let B denote the event that you are shown that a goat is behind door #3. The probability that there is a goat behind door #3 is best calculated by considering three conditional probabilities.

The probability that you are shown the goat is behind door #3 is an impossible event is the car is behind door #3. Thus $\Pr(B|A_3) = 0$. Since you picked door #1, Monty will open either door #2 or door #3, but not door #1. Thus if the car is actually behind door #2, it is a certainty that Monty will open door #3 and show you a goat. Thus $\Pr(B|A_2) = 1$. If you have picked correctly and have chosen the right door, then there are goats behind both door #2 and door #3. In this case, there is a 50% chance that Monty will open door #2 and a 50% chance that he will open door #3. Thus $\Pr(B|A_1) = 1/2$.

By Baye's theorem,

$$\boldsymbol{PPPP}(\boldsymbol{AA}_{11}\ |\boldsymbol{BB}) = \frac{\boldsymbol{PPP}\ \boldsymbol{P}\ (\boldsymbol{AA}_{11})\ \boldsymbol{PPPP}(\boldsymbol{BB}|\boldsymbol{AA}_{11})}{\boldsymbol{PPPP}(\boldsymbol{AA}_{11})\boldsymbol{PPPP}(\boldsymbol{BB}|\boldsymbol{AA}_{11}) + \boldsymbol{PPPP}(\boldsymbol{AA}_{00})\boldsymbol{PPPP}(\boldsymbol{BB}|\boldsymbol{AA}_{00}) + \boldsymbol{PPPP}(\boldsymbol{AA}_{33})\boldsymbol{PPPP}(\boldsymbol{BB}|\boldsymbol{AA}_{33})}$$

Plugging in the probabilities that we have derived, we find that

$$\boldsymbol{PPPP}(\boldsymbol{AA}_{11}|\boldsymbol{BB}) = \frac{(\mathbf{11/33})(\mathbf{11/00})}{(11/33)(11/00) + (11/33)(11) + (11/33)(00)} = \frac{\mathbf{11/66}}{11/66 + 11/33} = 11/33$$

Also,

$$\boldsymbol{PPPP}(\boldsymbol{AA}_{00}|\boldsymbol{BB}) = \frac{(\mathbf{11/33})(\mathbf{11})}{(11/33)(11/00) + (11/33)(11) + (11/33)(00)} = \frac{\mathbf{11/33}}{11/66 + 11/33} = 00/33$$

And since you already know that the car is not behind door #3, $\boldsymbol{PPPP}(\boldsymbol{AA}_{33}|\boldsymbol{BB}) = \mathbf{00}$.

Thus you have a 1/3 of picking the car if you stick with you initial choice of door #1, but a 2/3 chance of picking the car if you switch doors. It is in your interest to switch doors.

Did you think that there was no advantage to switching doors? You're not alone. Marilyn Vos Savant, famous as having the world's highest IQ at 228, wrote a column for *Parade* magazine for many years. In 1990 a reader posed the Monty Hall problem to her, and she provided the correct answer. But many people, including people with Ph.D.s, including some mathematicians, derided Marilyn for being wrong. Even the famous mathematician Paul Erdos found the problem to be counterintuitive (Hofmann, 1998). But the correct answer is that once door #3 is opened and revealed to have a goat behind it, there is a two-thirds chance that the car is behind door #2. If you're still not convinced, conduct a Monte Carlo simulation to see that this is the correct answer.

For our application of Bayes' Theorem to cost estimating we will need the continuous form of Baye's Theorem. If the prior distribution is continuous, Bayes' Theorem is written as

$$\pi\pi(\theta\theta|xx_{11}, \dots, xx_{nn}) = \frac{\pi\pi(\theta\theta)ff(xx_{11}, \dots, xx_{nn}|\theta\theta)}{ff(xx_{11}, \dots, xx_{nn})} = \frac{\pi\pi(\theta\theta)ff(xx_{11}, \dots, xx_{nn}|\theta\theta)}{\int \pi\pi(\theta\theta)ff(xx_{11}, \dots, xx_{nn}|\theta\theta)dd\theta\theta}$$

where:

$\boldsymbol{\pi\pi(\theta\theta)}$ is the *prior density*, the initial density function for the parameters that varies in the model. It is possible to define an *improper prior density*, one which is nonnegative but whose integral is infinite;

$\boldsymbol{ff(xx|\theta\theta)}$ is the conditional probability density function of the model. It defines the model's probability given the parameter $\theta\theta$;

$\boldsymbol{ff(xx_{11}, \dots, xx_{nn}|\theta\theta)}$ is the conditional joint probability density function of the data given $\theta\theta$. Typically the observations are assumed to be independent given $\theta\theta$, and in this case,

$$ff(xx_{11}, \dots, xx_{nn}|\theta\theta) = \prod_{ii=11}^{mm} ff(xx_{ii}|\theta\theta)$$

and $\boldsymbol{ff(xx_{11}, \dots, xx_{nn})}$ is the unconditional joint density function of the data $\boldsymbol{xx_{11}, \dots, xx_{nn}}$. It is calculated from the conditional joint density function by integrating over the prior density function of $\theta\theta$:

$$ff(xx_{11}, \dots, xx_{nn}) = \int \pi\pi(\theta\theta)ff(xx_{11}, \dots, xx_{nn}|\theta\theta)dd\theta\theta$$

$\boldsymbol{\pi\pi(\theta\theta|xx_{11}, \dots, xx_{nn})}$ is the *posterior density function*, the revised density function for the parameter $\theta\theta$ based on the observations $xx_1, \dots, xx_{nn}$.

$\boldsymbol{ff(xx_{nn+11}|xx_{11}, \dots, xx_{nn})}$ is the *predictive density function*, the revised unconditional density based on the sample data. It is calculated by integrating the conditional probability density function over the posterior density of $\theta\theta$:

$$ff(xx_{nn+11}|xx_{11}, \dots, xx_{nn}) = \int ff(xx_{nn+11}|\theta\theta)\pi\pi(\theta\theta|xx_{11}, \dots, xx_{nn})dd\theta\theta$$

Bayesian Regression

In this section we consider ordinary least squares CERs of the form

$$YY = aa + bbbb + \varepsilon\varepsilon$$

This involves prior distributions about a and b, as well as $\varepsilon\varepsilon$.

For the application of Bayesian regression, we will write this in mean deviation form:

$$YY = \alpha\alpha_{x\bar{x}} + \beta\beta(bb - \text{[illegible]}) + \varepsilon\varepsilon$$

This form makes it easier to establish prior inputs, since it is easier to think of an average value for prior cost than it is for the intercept of the least-squares equation.

For nonlinear regression we will consider log-transformed versions of the power equation $YY = aabb^{bb}$.

While weight is not a true cost driver, it is an excellent proxy for program scope and for most systems is highly correlated with cost. We will use weight-based CERs in our examples.

We begin with an example of sanitized cost and weight data for commercial-like earth-orbiting satellites. The claim has been made that these satellites are much cheaper than the average satellite purchased using traditional government acquisition practices. While only a handful of satellites have been built using the streamlined approach, they are on average much cheaper than traditional programs. The difference is remarkable - $4,000 per pound vs. $18,000 per pound, a reduction of more than 75%.

What is the best way to model these more economical satellites? I have a data set of 72 Earth-orbiting robotic satellites, but if I use that alone, I likely will significantly overestimate the cost.

An alternative would be to develop a CER for just the new missions alone. We can fit a power equation to this data set. See Figure 2. The R^2 coefficient is 63%, which is not bad. so why not just use that? Because the number of data points is an order of magnitude smaller than recommended for the minimum number of data points in a classical regression.

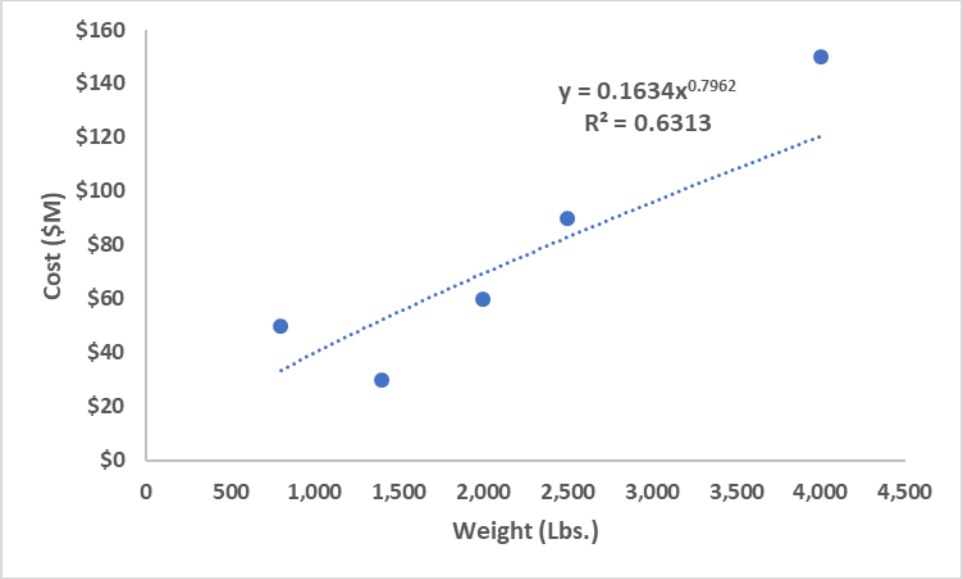

Figure 2. Regression of cost vs. weight for set of five data points.

In comparison to the larger set of 72 data points that do not include the five data points in the smaller set, we obtain another decent fit to the historical data, with an R^2 equal to 61%. See Figure 3.

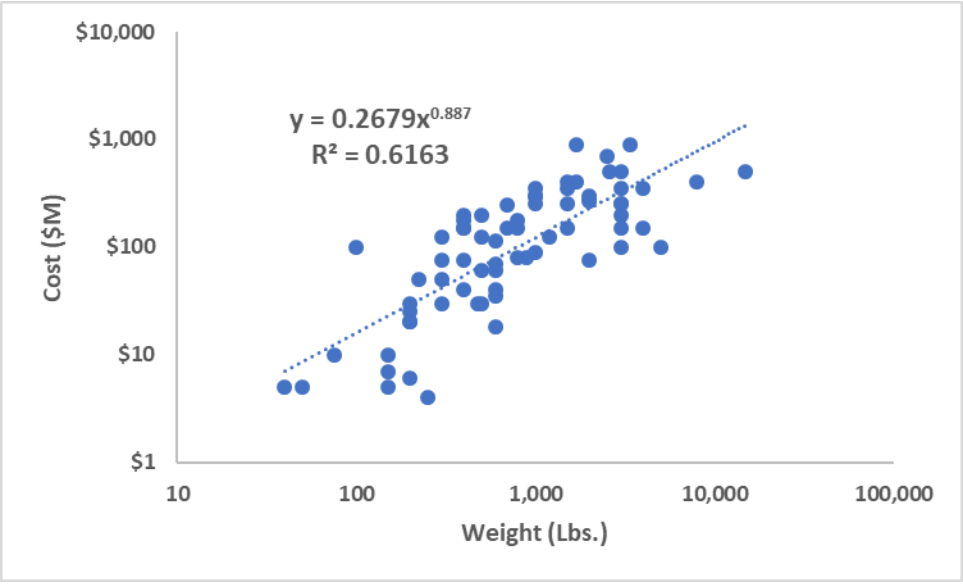

Figure 3. Regression of cost vs. weight for 72 earth-orbiting robotic satellites.

The Bayesian approach allows us to combine the earth-orbiting spacecraft data with the smaller data set. We use a specific type of hierarchical approach like the one used in credibility theory in insurance for setting premiums for property and casualty. We treat the earth-orbiting spacecraft data as our prior information. We update our prior information with the set of five data points. This specific type of hierarchical approach

involves two sets. One a general set, and the other a more specific subset or subtype. In this case, the commercial-like satellites are earth-orbiting satellites, and they are procured by government agencies. This is the general set. The commercial-like satellites are a proper subset of the earth-orbiting satellites. This is the specific subset.

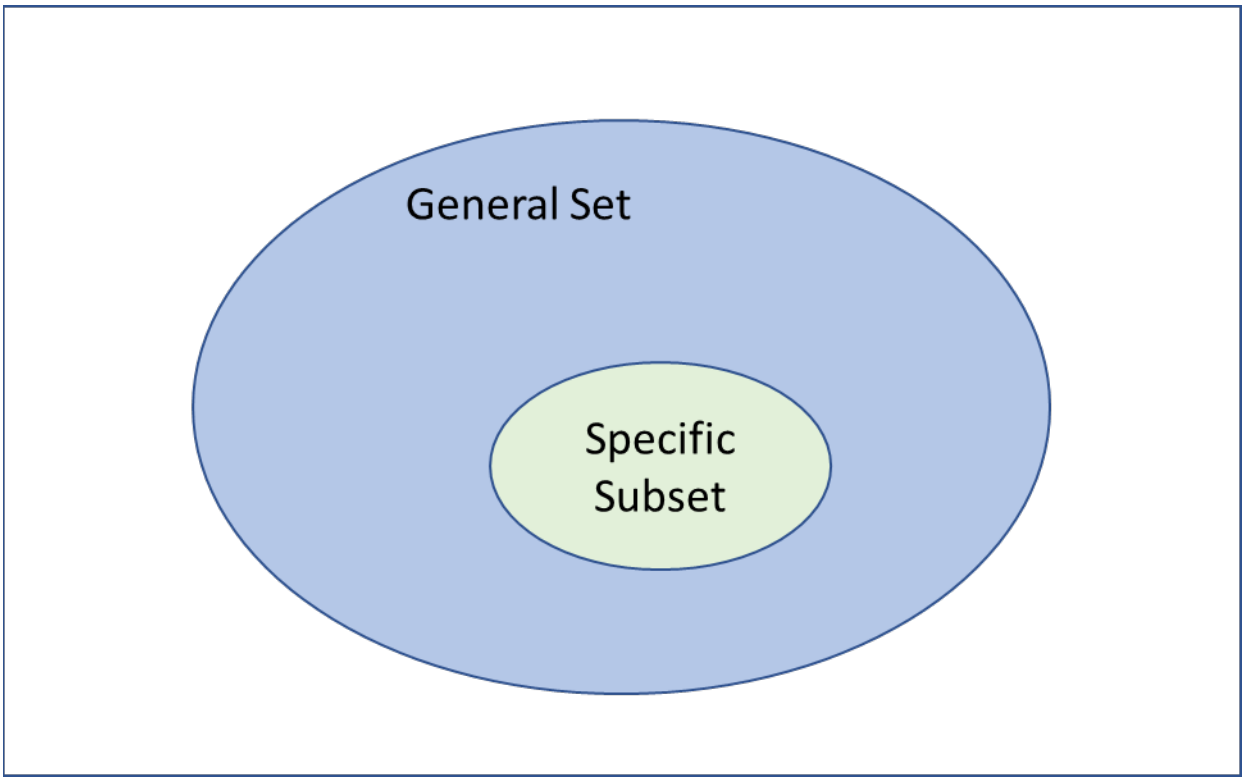

Figure 4. Conceptual illustration of the data used in a hierarchical approach.

This particular hierarchical approach has been used for over a century in setting property and casualty insurance premiums, particularly in smaller markets and for less common types of insurance.

There is a significant amount of probability theory in the derivation of Bayesian methods for regression. But I will avoid this here and refer the reader interested in the mathematical details to my earlier paper on this subject (Smart 2014).

The bottom line is that to implement the method you only need to understand the concept of standard error of the coefficients. To combine the two regression equations using Bayes' Theorem, we combine each coefficient, the intercept and the slope, separately. Bayes' Theorem combines the coefficients using the weighted averages of each individual coefficient. These weights are based on the amount of uncertainty relative to one another. The greater the uncertainty in the coefficient relative to the other coefficient, the smaller its weight. The smaller the uncertainty in the coefficient relative to the other coefficient, the greater its weight. This makes intuitive sense – we

should have greater confidence in the coefficient with the less variation, and assign it a greater weight in taking the average of the two.

The formula for combining the two coefficients is thus very simple. It takes the inverse of each of the variances, which is called the *precision*, and forms a weighted average of the precisions of each coefficient.

Given two independent, unbiased coefficients $\theta\theta\hat{}_1$ and $\theta\theta\hat{}_2$ with precisions $\rho\rho_1 = 1/VVaaPP\hat{}\theta\theta\hat{}_1\hat{}$ and $\rho\rho_2 = 1/VVaaPP\hat{}\theta\theta\hat{}_2\hat{}$ respectively, the minimum variance estimate of the weighted average is given by

$$\frac{\rho\rho_1}{\rho\rho_{11} + \rho\rho_{00}}\hat{}_{11} + \frac{\rho\rho_0}{\rho\rho_{11} + \rho\rho_{00}}\hat{}_{00}$$

The variance is the square of the standard error. The standard error of each coefficient is provided by statistical packages, including the data analysis add-in for Excel and the "LINEST" function in Excel.

We will next show how the weight average is calculated in Excel using the LINEST function. LINEST is a matrix function. To enter it you select a range of cells for the output, type in "=LINEST(y values, x values, true, true)." The first "true" is to include the intercept, and the second is to include all the statistics. The output for a single independent variable appears as in Table 1.

b-value	**a-value**
Std err (b)	**Std err (a)**
R^2	**Std Err of Est**
F Statistic	**Degrees of Freedom**
SS Reg	**SS Resid**

Table 1. "LINEST" Excel function output format.

The first row are the coefficients. The "a-value" is the intercept and the "b-value" is the coefficient of the slope. The second row is the standard errors of each coefficient. These are the only two sets of values we need from each equation to apply Bayes' Theorem.

Recall that we are using the form

$$YY = \alpha\alpha_{xx\bar{}} + \beta\beta(bb - \hat{}) + \varepsilon\varepsilon$$

and we are applying it to the log transformations of the data.

When we apply the LINEST function to the larger data set we obtain the results in Table 2.

0.8858	**4.6087**
0.0809	**0.0956**
0.6311	**0.8111**
119.7361	**70**
78.7810	**46.0569**

Table 2. LINEST results for larger data set.

The a-value coefficient is equal to *4.6087* and its standard error is *0.0956*. The b-value coefficient is *0.8858* and its standard error is *0.0809*.

See Table 3 for the LINEST results for the smaller data set.

0.8144	**4.1359**
0.2588	**0.1416**
0.7675	**0.3167**
9.9033	**3**
0.9930	**0.3008**

Table 3. LINEST results for the small data set.

The a-value coefficient for the smaller data set is equal to *4.1359* and its standard error is *0.1416*. The b-value coefficient is *0.8144* and its standard error is *0.2588*. Note that the standard error calculation involves division by the degrees of freedom, so there is a penalty applied for having less data.

To combine the a-value coefficient we calculate the variances of each as the square of the standard error. Thus the variance of the a-value for the larger data set is equal to $0.0956^2 \approx 0.0091$ and the variance of the a-value for the smaller data set is equal to $0.1416^2 \approx 0.0201$. The weight for the larger data set's a-value coefficient is calculated as

$$\frac{\frac{1}{0.0091}}{\frac{1}{0.0091}+\frac{1}{0.0201}} \approx 68.7\%$$

and the coefficient weight for the a-value of the smaller data set is *1-0.687 = 31.3%*.

For the b-values we use similar calculations to determine the coefficient weight for the larger data set as

$$\frac{\frac{1}{0.0065}}{\frac{1}{0.0065}+\frac{1}{0.0670}} \approx 91.1\%$$

and the coefficient weight for the b-value of the smaller data set is *1-0.911 = 8.9%*.

The Bayesian estimate of the a-value coefficient is thus *4.6087*0.687+4.1359*.313 = 4.4607*, and the b-value coefficient is *0.8858*0.911+0.8144*.089 = 0.8794*.

We have used two different data sets, each of which has its own average weight, but since we consider the smaller data set as our "sample" and the larger data set as "prior

information" we use the mean of the logs of the weights for the smaller data set, which yields

$$llll(CCCCCCCC) = 4.4607 + 0.8794(llll(WWWWWWWWhCC) - 7.5161) + \varepsilon\varepsilon$$

Rearranging terms results in

$$llll(CCCCCCCC) = -2.1491 + 0.8794llll(WWWWWWWWhCC) + \varepsilon\varepsilon$$

Exponentiating yields the power equation

$$CCCCCCCC = 0.1166WWWWWWWhCC^{0.8794} \cdot \varepsilon\varepsilon$$

In all equations the error term denoted by the Greek letter epsilon simply denotes the difference between the estimated cost and the actual.

Note that because of the equation form we use to average the two data sets using Bayes' Theorem, the a-value coefficient in this case happens to be smaller than either coefficient of the two regressions for the individual data sets. Comparing the estimates provided by these equations for another commercial-like acquisition whose actual weight was 3,280 lbs. and whose cost was $180 million, we find that the equation based on the smaller data set alone predicts cost at $100 million, while the equation based on the larger data set alone predicts cost at $368 million. The Bayesian regression equation provides an estimate that is closer than either of these to the actual, at $144 million.

This particular approach to Bayesian estimating with a normal prior, normal likelihood, and known variance has also been used in software cost estimating (Boehm et al., 2018). In that particular application the prior is established via expert judgment.

Assumptions

There are two simplifying assumptions made that are dubious in this type of analysis.

One is the assumption that the variance of the estimating equation is known. That is, not only is it a fixed quantity, but we know the exact value. However, the value that we use is estimated from the likelihood, which is the smaller data set. Thus, we have a high degree of uncertainty in the variance, which means this assumption is not valid.

A second assumption is that the smaller data set has residuals that are normally distributed. The assumption that the residuals are normally distributed in the larger data set is not an issue, however when you have a small data set with an unknown variance, the residuals are better modeled with a Student's *t* distribution. The use of the Student's t distribution to model the log-space residuals of log-transformed ordinary least squares was first advocated several years ago at an ISPA-SCEA conference (Druker et al. 2009), and has since been implemented in the ACE-IT cost estimating platform.

The distribution of the variance is easy to estimate, and we will deal with that next. Incorporating the change from known variance and the use of the Student's *t* distribution to model the residuals is more challenging, and we will deal with that second.

Distribution of the unknown variance

By Cochran's Theorem (Cochran 1934), when we have *p* parameters in a linear regression,

$$\frac{\sum_{ii=1}^{nn} \left(YY_{i} - \hat{YY}_{ii}\right)^{2}}{\sigma\sigma^{2}} \sim \chi\chi^{2}(ll - pp)$$

where $\chi\chi^{2}(ll - pp)$ is a chi-square distribution with *n-p* degrees of freedom.

Also, we estimate the square of the standard error as

$$\hat{\sigma}^{2} = \frac{\sum_{ii=1}^{nn} \left(YY_{i} - \hat{YY}_{ii}\right)^{2}}{ll - pp}$$

So we have that

$$\frac{(ll - pp)\hat{\sigma}^{2}}{\sigma\sigma^{2}} \sim \chi\chi^{2}(ll - pp)$$

We are interested in the distribution of $\sigma\sigma^{2}$, which follows a scaled-inverse chi-square distribution with parameters *n-p* and $\hat{\sigma}^{2}$.

The scaled-inverse chi-square distribution has probability density function given by

$$\frac{\left(\frac{ll - pp}{2}\right)^{\frac{nn-pp}{2}} \hat{\sigma}^{2\frac{nn-pp}{2}} WW^{-\frac{nn-pp}{2}} e^{-\frac{nn-pp}{2}\frac{\hat{\sigma}^{2}}{\sigma\sigma^{2}}}}{\Gamma\left(\frac{ll - pp}{2}\right) \sigma\sigma^{2\left(1+\frac{nn-pp}{2}\right)}}$$

The scaled-inverse chi-square is only defined over positive values and has positive skew. See Figure 1 for a graph of a scaled-inverse chi-square with parameters *n-p* = 70 and $\hat{\sigma}^{2} = 66\%$.

Also the variance follows an inverse gamma with parameters $\frac{nn-pp}{2}$ and $\frac{nn-pp}{2}\hat{\sigma}^{2}$, and the inverse of the variance follows a gamma distribution with the same parameters. The inverse gamma and gamma distributions are more commonly encountered than the scaled-inverse chi-square and is more commonly available in statistical software.

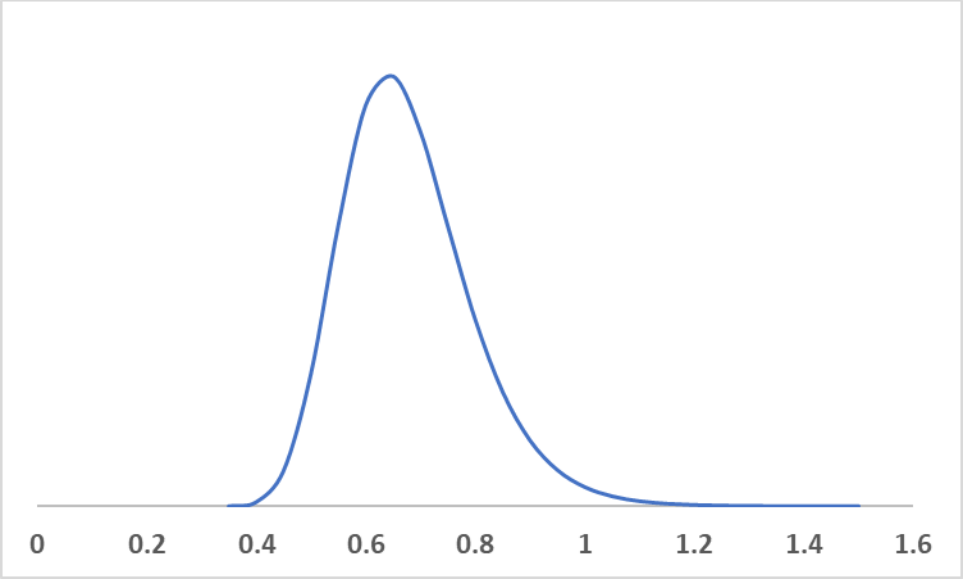

Figure 5. Probability density function of a scaled-inverse chi-square with parameters 70 and 0.66.

Conjugate priors

One of the nice features of using the normal prior and normal likelihood model with known variance is that the posterior is also a normal distribution. That is, the prior and the posterior have the same distribution. When the prior and posterior have the same distribution using conjugate priors makes calculating the posterior relatively easy and analytically tractable. Once you move away from conjugate priors you lose analytical tractability.

Once you remove the assumption that the variance is known, we lose the relatively straightforward normal conjugate prior model for regression. Recall that the essence of Bayes' theorem is that the posterior is proportional to the product of the likelihood and the prior distribution, that is for uncertain parameters y and θ,

$$pp(\theta\theta|yy) \propto pp(\theta\theta)pp(yy|\theta\theta)$$

When the regression is centered about zero (as described in the previous section), the variance is known and the prior and likelihood are both normal, then the likelihood is (note that in the remainder of the paper we use bold font to denote matrices and vectors)

$$pp(\boldsymbol{yy}|\boldsymbol{XX}, \boldsymbol{\beta\beta}, \mu\mu, \varpi^2) = \boldsymbol{NN}(\boldsymbol{yy}|\boldsymbol{XX}, \boldsymbol{\beta\beta}, \sigma\sigma^2) \propto \underset{WW}{} \quad -\frac{1}{2\sigma\sigma^2} |\boldsymbol{\hat{y}yy} - \boldsymbol{\hat{y}\hat{y}\hat{y}} - \boldsymbol{XX\beta\beta\hat{y}}|^{00}_{0}$$

Where $\hat{y}$ is the sample average of the dependent variable; the joint prior of the coefficients is

$$pp(\boldsymbol{\beta\beta}) = NN(\boldsymbol{\beta\beta}|\boldsymbol{\beta\beta}_{00}, \boldsymbol{VV}_{00})$$

Then by Bayes' Theorem, the posterior is

$$pp(\boldsymbol{\beta\beta}|\boldsymbol{XX},\ \boldsymbol{yy},\ \sigma\sigma^2)\ \propto\ NN(\boldsymbol{\beta\beta}|\boldsymbol{\beta\beta}_{00},\ \boldsymbol{VV}_{00})\ \cdot\ NN(\boldsymbol{yy}|\boldsymbol{XX\beta\beta},\ \sigma\sigma^2\boldsymbol{II}_{nn})\sim NN(\boldsymbol{\beta\beta}|\boldsymbol{\beta\beta}_{NN},\ \boldsymbol{VV}_{NN})$$

where

$$\boldsymbol{\beta\beta}_{NN} = \boldsymbol{W}_{NN}\boldsymbol{VV}_0^{-11}\boldsymbol{\beta\beta}_{00} + \frac{1}{\sigma\sigma^2}\boldsymbol{VV}_{NN}\boldsymbol{XX}^{TT}\boldsymbol{yy}$$

$$\boldsymbol{VV}_{NN}^{-11} = \boldsymbol{VV}_{00}^{-11} + \frac{1}{\sigma\sigma^2}\boldsymbol{XX}^{TT}\boldsymbol{XX}$$

$$\boldsymbol{VV}_{NN} = \sigma\sigma^2\left(\sigma\sigma^2\boldsymbol{VV}_0^{-11} + \boldsymbol{XX}^{TT}\boldsymbol{XX}\right)^{-11}$$

See Appendix 1 for the derivation.

The posterior predictive distribution is also normal, since

$$pp(\boldsymbol{yy}|\boldsymbol{xx},\ \mathcal{DD},\ \sigma\sigma^2)\ =\ \int NN(\boldsymbol{yy}|\boldsymbol{xx}^{TT}\boldsymbol{\beta\beta},\ \sigma\sigma^2)NN(\boldsymbol{\beta\beta}|\boldsymbol{\beta\beta}_{NN},\ \boldsymbol{VV}_{NN})d\boldsymbol{\beta\beta}\ =\ NN(\boldsymbol{yy}|\boldsymbol{xx}^{TT}\boldsymbol{\beta\beta}_{NN},\ \sigma\sigma^2\ +\ \boldsymbol{xx}^{TT}\boldsymbol{VV}_{NN}\boldsymbol{xx})$$

Note that we don't have to calculate the integral to derive the predictive posterior, we just need to note that since $\boldsymbol{yy} = \boldsymbol{xx}^{TT}\boldsymbol{\beta\beta} + \boldsymbol{\varepsilon\varepsilon}_{11}$ where $\varepsilon\varepsilon_1 \sim NN(\boldsymbol{00}, \boldsymbol{\sigma\sigma}^{00})$ and $\boldsymbol{\beta\beta} = \boldsymbol{xx}^{TT}\boldsymbol{\beta\beta}_{NN} + \boldsymbol{\varepsilon\varepsilon}_{00}$ where $\boldsymbol{\varepsilon\varepsilon}_{00} \sim NN(0, \boldsymbol{VV}_{NN})$, then $\boldsymbol{yy} = \boldsymbol{xx}^{TT}(\boldsymbol{\beta\beta}_{NN} + \varepsilon\varepsilon_2) + \varepsilon\varepsilon_1 = \boldsymbol{xx}^{TT}\boldsymbol{\beta\beta}_{NN} + \boldsymbol{xx}^{TT}\boldsymbol{\varepsilon\varepsilon}_2 + \varepsilon\varepsilon_1 \sim NN(\boldsymbol{xx}^{TT}\boldsymbol{\beta\beta}_{NN}, \sigma\sigma^2 + \boldsymbol{xx}^{TT}\boldsymbol{VV}_{NN}\boldsymbol{xx})$.

The mean of the predictive distribution is thus the mean of the posterior, and the variance is the sum of the observed variance and another term that depends on the variance of the parameters of the posterior distribution. This latter value depends on the degree to which the input value for the prediction is close to the training data.

In the case of unknown variance it turns out that if we use the inverse gamma distribution to model the variance, as discussed in the previous section, then with normal prior on the coefficients conditional on the variance, and normal likelihood, we have that

$$pp(\boldsymbol{\beta\beta}, \sigma\sigma^2)\ =\ pp(\boldsymbol{\beta\beta}|\sigma\sigma^2)pp(\sigma\sigma^2)$$

In this case $pp(\boldsymbol{\beta\beta}|\sigma\sigma^2)$ follows a normal distribution, and $pp(\sigma\sigma^2)$ follows an inverse gamma distribution. This conditional product is called a normal-inverse gamma distribution, that is,

$$pp(\boldsymbol{\beta\beta},\ \sigma\sigma^2)\ =\ pp(\boldsymbol{\beta\beta}|\sigma\sigma^2)pp(\sigma\sigma^2)\sim NN(\boldsymbol{\beta\beta}|\boldsymbol{\beta\beta}_{00},\ \sigma\sigma^2\boldsymbol{VV}_{00})IIII(\sigma\sigma^2|aa_0,\ bb_0)$$

$$= \frac{bb_0^{aa_0}}{(2\pi\pi)^{\frac{D}{2}}|\boldsymbol{VV}_{00}|^{0.5}\Gamma(aa_0)}(\sigma\sigma^2)^{-(aa_0+\frac{DD}{2}+1)}\cdot \exp\left(-\frac{1}{2\sigma\sigma^2}(\boldsymbol{\beta\beta}-\boldsymbol{\beta\beta}_{00})^{TT}\boldsymbol{VV}_{00}^{-11}(\boldsymbol{\beta\beta}-\boldsymbol{\beta\beta}_{00}) + 2bb_0\right)$$

Where D is the number of variables in the regression equation. With the normal-inverse gamma prior and a normal likelihood, the posterior joint distribution of the coefficients and the variance is also a normal-inverse gamma with

$$pp(\boldsymbol{\beta\beta},\ \sigma\sigma^2|\boldsymbol{XX},\ \boldsymbol{yy})\ =\ NN(\boldsymbol{\beta\beta}|\boldsymbol{\beta\beta}_{NN},\ \sigma\sigma^2\boldsymbol{VV}_{NN})IIII(\sigma\sigma^2|aa_{NN},\ bb_{NN})$$

where

$$\boldsymbol{\beta\beta}_{NN} = \boldsymbol{VV}_{NN}(\boldsymbol{VV}_0^{-1}\boldsymbol{\beta\beta}_0 + \boldsymbol{XX}^{TT}\boldsymbol{yy})$$

$$\boldsymbol{VV}_{NN} = (\boldsymbol{VV}_0^{-1} + \boldsymbol{XX}^{TT}\boldsymbol{XX})^{-1}$$

$$aa_{NN} = aa_0 + ll/2$$

$$bb_{NN} = bb_0 + \frac{1}{2}(\boldsymbol{\beta\beta}^{TT}\boldsymbol{VV}^{-1}\boldsymbol{\beta}_0\boldsymbol{\beta}_0^{TT} + \boldsymbol{yy}^{TT}\boldsymbol{yy} - \boldsymbol{\beta\beta}_{NN}^{TT}\boldsymbol{VV}_{NN}^{-1}\boldsymbol{\beta\beta}_{NN})$$

See Appendix 2 for the derivation of this last expression.

The posterior predictive distribution in this case is not a normal distribution. Rather it is a Student's *t* distribution, i.e.,

$$pp(\boldsymbol{yy}|\boldsymbol{xx}, \mathcal{DD}) = \mathcal{TT}(\boldsymbol{yy}|\boldsymbol{xx}\boldsymbol{\beta\beta}_{NN}, \frac{bb_{NN}}{aa_{NN}}(1 + \boldsymbol{xx}^{TT}\boldsymbol{VV}_{NN}\boldsymbol{xx}), 2aa_{NN})$$

With mean = $\boldsymbol{XX}\boldsymbol{\beta\beta}_{NN}$, variance = $\frac{bb_{NN}}{aa_{NN}}(1 + \boldsymbol{XX}^{TT}\boldsymbol{VV}_{NN}\boldsymbol{XX}^{TT})$, and degrees of freedom = $2aa_{NN}$

See Appendix 3 for the derivation.

For our particular application, the number of degrees of freedom for the *t* distribution is roughly the sum of the number of data points in the prior and the number of data points in the likelihood. As long as the total of these two is at least 30, the Student's *t* distribution will be approximately normal.

The expression for the coefficients and the variance matrix is like the case when the variance is known – the difference is that the variance disappears from the two expressions once we remove the assumption that it is known.

Markov Chain Monte Carlo

When we relax the assumption of known variance, the result is still analytically tractable, if the likelihood is normal. What happens if we relax this assumption and use a Student's t distribution instead? In this case - a normal prior and a Student's t likelihood – the resulting posterior is not a conjugate prior so the result cannot be derived analytically. In this case we must turn to a simulation method. Most cost analysts are familiar with Monte Carlo simulation. Markov chain Monte Carlo simulation is a specific type of Monte Carlo simulation that is different than what is typically used in cost risk analysis. To begin, Markov chain Monte Carlo simulation picks a random parameter value to consider. The simulation will continue to generate random values subject to some rule for determining what makes a good parameter value. Given the prior distribution, if the simulation generated parameter value is better than the previous at explaining the data, it is added to the chain of parameter values with a probability determined by how much better it is than the previous.

The basic idea in Markov chain Monte Carlo is that we sample each variable in turn, conditioned on the values of all the other variables in the distribution. Given a joint sample *x(n),* we generate a new sample *x(n+1)* by sampling each component in turn, based on the values of *x(n)*. If we have two variables, x_1 and x_2, we calculate $x_1(n+1) \sim p(x_1|x_2(n))$ and $x_2(n+1) \sim p(x_2|x_1(n+1))$.

As mentioned earlier, we start the simulation from an arbitrary state. Because of the conditional simulations, it takes some time for the chain to converge to its stationary distribution. The samples collected from the period before the chain converges are discarded. These discarded samples are generated during what is called the burn-in phase. Because we start from an arbitrary state, it is common to use multiple initial values, which generates multiple chains, one for each set of initial values for the parameters.

We discuss two tools that can be used to do the Markov chain Monte Carlo simulations. One is R – we can use the built-in statistical capabilities in base R to conduct the simulation. Another is WinBUGS, one of several simulation tools that are designed specifically for doing a particular type of Markov chain Monte Carlo called Gibbs sampling. WinBUGS is free Windows software for doing Bayesian inference Using Gibbs Sampling. WinBUGS requires some simple programming. The syntax is similar to R, but because it is specifically design to solve this particular kind of problem, less code is required than in base R. WinBUGS also has numerous diagnostic tools and graphs. However, if you want to do analysis, you need to export the simulations results back to R. One option for working around this is to use the R2OpenBUGS package for R to run WinBugs from R, which then produces the results in R for additional analysis. We provide the R code for the case of the student t likelihood with unknown variance.

Note that in R, comment lines begin with the hashtag symbol #. In the R code, we set the number of trials equal to 10,000, with a burn-in period equal to the first 1,000 trials.

In the Markov chain Monte Carlo simulation, we calculate the likelihood of the posterior for the candidate values. If the difference between the loglikelihood of the current values is larger than a random value, we accept. Otherwise we reject. We reject some of the values with higher likelihood, and accept some with lower likelihood, in order to more fully explore the parameter space. Accepting too many candidate samples leads to highly correlated draws over time. Thus we want a relatively large rejection rate, on the order of 40-80%.

The R code is simple. You don't have to actually code the simulation, that work is done for you. You merely must specify the likelihood, the priors, the data, and the initial values.

R code:

```
install.packages("R2OpenBUGS")

library('R2OpenBUGS')
```

```
#Example
 x = c(8.2938,7.750,7.2235,6.6598,7.6533,8.0956)
 y = c(4.9182,4.5276,3.5965,3.68437,3.9526,NA)
 n <- length(x)
 data <- list("x","y","n")
 inits <- function(){
  list(beta=c(5,0.7),tau=1,gamma=c(0,0,0,0,0,0))
  list(beta=c(4,0.9),tau=0.1,gamma=c(0,0,0,0,0,0))
 }
 Case3 = bugs(data,inits,
              model.file="model.txt",
              parameters=c("beta","tau","gamma"),
              n.chains=1,n.iter=20000,n.burnin=5000,n.thin=1,
              codaPkg=FALSE, debug = TRUE)
summary(Case3)
```

We compare the results of the three cases in Table 4. Case 1 is the normal prior and normal likelihood with known variance. Case 2 is the normal prior and normal likelihood with unknown variance. Case 3 is the normal prior and Student's t likelihood with unknown variance.

	Case 1	Case 2	Case 3
log-space intercept	4.4607	4.5800	4.5780
linear intercept	0.8794	0.8846	0.8880
slope	0.1166	0.1263	0.1229
prediction ($ millions)	144	163	163

Table 4. Comparison of the results of the three cases.

We see that the mean results of cases 2 and 3 are virtually identical. We recognize that changing the assumption of the variance causes less weight to be placed on the likelihood, and more on the prior. Recalling that the actual cost is equal to $180, we see that removing the assumption of known variance results in a more accurate prediction.

What about the variance? Before we delve into that, note that Bayesian analysis produces Bayesian credible intervals rather than frequentist prediction intervals.

Traditional frequentist prediction intervals are not very intuitive. In the long run, with data from numerous samples, a 95% prediction interval calculated from each sample will contain the true parameter 95% of the time. By contrast, a 95% Bayesian credible interval contains the true parameter value contains the true parameter value with 95% probability.

We are interested in the variance because we need not only the mean, but the distribution type and the variance to model the uncertainty and risk for the estimate. For case 1, the predictive variance at the new estimate in log space is *0.1085*, only slightly higher than the variance of the likelihood. For case 2, when we relax the assumption that the variance is known, the predictive distribution is a Student's t distribution with *75* degrees of freedom and variance equal to *0.9061*, much higher than for case 1, but also much more realistic. For case 3, there is no analytical predictive distribution so we have to simulate it from the results of the Markov chain Monte Carlo simulation. The R code for this is (considered in the context of the previous code):

```
Ynew<-numeric(T)

munew<-numeric(T)

for(j in B+1:T){

  munew[j]<-alph[j]+beta[j]*(8.0956-7.5161)

  Ynew[j]<-rt(1,3)*(sigma2[j])^0.5+munew[j]

}
```

In this code, *8.0956* is the new weight in log space, and *7.5161* is the average log space weight for the sample.

The sample variance is *2.1508* in log space. The histogram is very wide, and ranges from *-31* to *84*. A truncated histogram from *0* to *15* is displayed in Figure 5. Fitting a normal distribution using the MASS (Modern Applied Statistics with S) package in R, we fit a normal and a Student's t distribution. The normal distribution has variance equal to the sample variance, while the Student's t has variance equal to *2.15084* and three degrees of freedom.

Note that a normal distribution in log space is a lognormal distribution in unit space, by definition. A Student's t distribution in log space is a log-t distribution in unit space. A

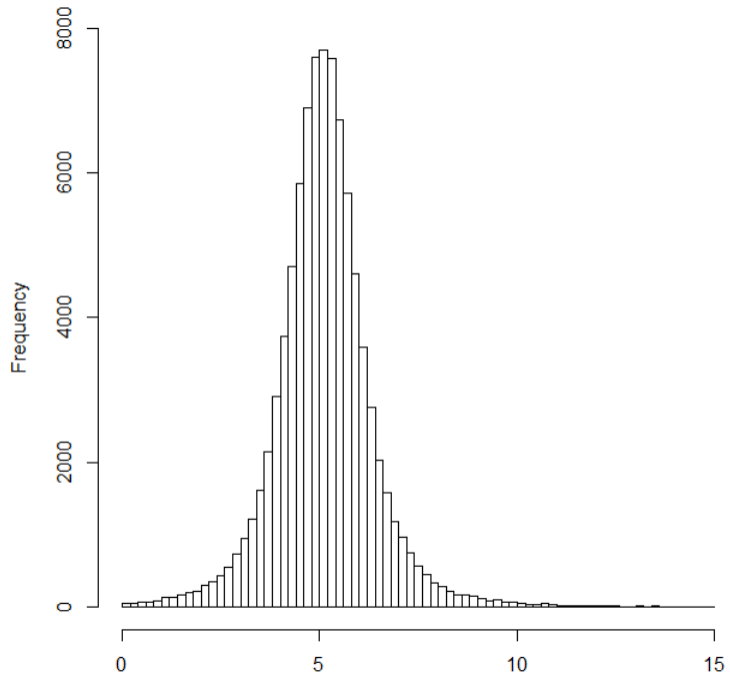

Figure 6. Truncated histogram for Student's t likelihood and unknown variance.

lognormal distribution has finite mean and variance, and you can parameterize a lognormal by these two parameters. A log-t distribution has neither a finite mean nor a finite variance. As a result, it has an extremely heavy right tail. See Table 5 for a comparison of the posterior predictive S-curves generated by the four different assumptions. The first assumption, normal likelihood of the estimate (in log space) with known variance results in the narrowest S-curve, with a 5^{th}-95^{th} percentile range equal to $\$164$ million. Relaxing the variance assumptions leads to a wider S-curve, with a 5^{th}-95^{th} percentile range equal to $\$762$ million. Changing the distribution of the likelihood to a student's t in log space results in two different S-curves, depending on whether a normal distribution or a t-distribution is used to model the predictive distribution in log space, which is a lognormal or log-t distribution in unit space. A lognormal distribution has a 5^{th}-95^{th} percentile range equal to $\$1.8$ billion. The log-t distribution has a narrower range from the 5^{th} to the 95^{th}, equal to $\$1.1$ billion. However, the tail of the log-t is very heavy which causes the range to be narrower for the log-t since more of the weight is in the far right tail, which is beyond the 95^{th} percentile.

Confidence level	Normal likelihood with known variance	Normal likelihood with unknown variance	Student t likelihood, unknown variance, normal error	Student t likelihood, unknown variance, Student's t error
5%	$84	$33	$15	$23
10%	$94	$48	$25	$41
20%	$109	$73	$47	$72
30%	$121	$99	$76	$100
40%	$132	$128	$112	$129
50%	$144	$163	$163	$163
60%	$156	$207	$236	$205
70%	$171	$269	$352	$266
80%	$190	$365	$560	$370
90%	$220	$558	$1,068	$642
95%	$247	$795	$1,819	$1,169
99%	$310	$1,565	$4,942	$7,295
99.5%	$336	$2,016	$7,125	$21,662
99.9%	$398	$3,434	$15,151	$842,898

Table 5. S-curve comparison for the different assumptions. Costs are in millions of dollars.

Although the log-t range is narrower, once you go to the *99th* percentile and above, the log-t distribution begins to blow up, going from $*7* billion at the *99*th percentile to $*843* billion at the *99.9th* percentile. That is, there is a *1* in *1,000* chance that the cost will at or above $*843* billion for an estimate whose *50th* percentile is $*163* million. To get a more succinct sense of how these S-curves compare, we look at the coefficients of variation (CV), which is the ratio of the standard deviation of the estimate to its mean. A variety of cost growth studies indicate that a reasonable value for CV is *50%* at the beginning of development (Smart 2018). The CV for the base case with normal likelihood and known variance is *34%*, lower than the rule of thumb. The CV for normal likelihood and unknown variance is *94%*, much higher and definitely in the reasonable range. When we change the likelihood to a Student's t-distribution, the CV when we fit a normal in log space (lognormal in unit space) to the predictive equation is *275%*; when we fit a Student's t-distribution to the predictive equation, this is a log-t in unit space, which since it does not have a finite mean or variance, does not exist.

Even though the assumption of a t-likelihood is more correct in the case of small samples, we need to use some common sense in establishing S-curves. A risk range that includes values orders of magnitudes above the median is not reasonable. On the other extreme, the assumption of known variance is unrealistic, so a compromise that assumes unknown variance but normal likelihood seems reasonable. Since we are leveraging our experience from a larger data set I believe this makes sense.

Summary

In this paper we have extended the linear model by changing two key assumptions. One of these is the assumption of known variance, and the second is the assumption of a normal distribution about the likelihood of the log-transformed ordinary least squares equation.

The assumption of known variance is not at all reasonable. We do not know the variance. What is worse is that the "known" variance is actually estimated from the likelihood. The likelihood in our problem set up has a small sample size, so even the estimate of this variance has significant uncertainty. We showed that the variance follows a specific distribution, namely an scaled inverse chi-squared (or equivalently an inverse gamma), and we can estimate the parameters for the prior from the number of degrees of freedom and the estimate variance of the estimating equation in the regression analysis used to establish the prior distribution.

The assumption of a normal likelihood is suspect. The problem set up is designed to work with likelihoods for which we have a small number of data points, otherwise we would have applied a traditional frequentist regression method. In this case, as has been pointed out by others (Druker et al., 2009), in such situations a Student's t distribution is more appropriate.

We provided an analysis of the example by relaxing the assumptions. First we relaxed the assumption of known variance, which leads to a tractable result. The inclusion of uncertainty about the variance of the likelihood in the form of a prior distribution for the estimating variance leads to greater weight on the prior and less weight on the likelihood, which changes the mean. The posterior predictive equation follows a Student's t distribution with high degrees of freedom as it benefits from the large sample used to establish the prior. For this number of degrees of freedom, the normal and Student's t distribution are indistinguishable so we can model the predictive equation with a normal distribution.

Second, we continue with unknown variance, but now we change the assumption that the likelihood is normally distribution to a Student's t distribution. This is not analytically tractable since we have departed from the realm of conjugate priors. In this case we have to resort to a specific type of Monte Carlo simulation called Markov chain Monte Carlo. We showed how this can be accomplished using both a statistical programming platform called R, and a platform specifically designed for this type of Bayesian analysis called WinBUGS.

The assumption of known variance drives the mean. Changing to unknown variance significantly increases the mean and noticeably increases the variance. Changing this in turn to a Student's t likelihood does not change the mean but drives the variance much higher.

In conclusion, using the Bayesian approach can benefit a cost estimator when they are faced with an all-too common situation of limited data. We have extended the simpled linear model to include more realistic assumptions. However, when departing from the

assumption that the likelihood is normally distributed, common sense should be applied when applying an extremely heavy-tailed distribution. The approaches used in this model also apply when the prior is specified using expert judgment.

Next Steps

This paper deals with parametric models, ones which use specific distributions, such as the normal, lognormal, or Student's t-distribution. It does not cover regression methods that do not use a distribution to model the residuals, such as the Minimum Unbiased Percentage Error (MUPE) method or the Zero-bias Minimum Percent Error (ZMPE) method. Markov Chain Monte Carlo simulation can be applied to histograms for both the prior and likelihood, and a future paper will deal with the application of MCMC techniques to ZMPE and MUPE regression methods.

Appendix 1

In both this appendix and in Appendix 2, we make use of the matrix identity

$$\boldsymbol{uu}^{TT}\boldsymbol{AAuu} - 2\boldsymbol{\alpha\alpha}^{TT}\boldsymbol{uu} + \boldsymbol{\alpha\alpha}^{TT}\boldsymbol{AA}^{-1}\boldsymbol{\alpha\alpha} = (\boldsymbol{uu} - \boldsymbol{AA}^{-1}\boldsymbol{\alpha\alpha})^{TT}\boldsymbol{AA}(\boldsymbol{uu} - \boldsymbol{AA}^{-1}\boldsymbol{\alpha\alpha})$$

where $\boldsymbol{u}$ and $\boldsymbol{\alpha}$ are vectors, and $\boldsymbol{A}$ is a matrix.

We need to combine the two exponents, and have it result in something that is quadratic in $\boldsymbol{\beta}$. Thus, we want the $\boldsymbol{u}$ term in the matrix identity to represent $\boldsymbol{\beta}$.

We set $\boldsymbol{\alpha\alpha} = \boldsymbol{VV}_0^{-1}\boldsymbol{\beta\beta}_0 + \frac{1}{\sigma\sigma^2}\boldsymbol{XX}^{TT}\boldsymbol{yy}$ and $\boldsymbol{AA} = \boldsymbol{VV}_0^{-1} + \frac{1}{\sigma\sigma^2}\boldsymbol{XX}^{TT}\boldsymbol{XX}$.

We begin with the expression $(\boldsymbol{\beta\beta} - \beta\beta_0)^{TT}\boldsymbol{VV}_0^{-1}(\boldsymbol{\beta\beta} - \beta\beta_0) + \frac{1}{\sigma\sigma^2}(\boldsymbol{yy} - \boldsymbol{XX\beta\beta})^{TT}(\boldsymbol{yy} - \boldsymbol{XX\beta\beta})$.

Expanding this we find

$$\boldsymbol{\beta\beta}^{TT}\boldsymbol{VV}_0^{-1}\boldsymbol{\beta\beta} - 2(\boldsymbol{VV}_0^{-1}\boldsymbol{\beta\beta}_0)^{TT}\boldsymbol{\beta\beta} + \boldsymbol{\beta\beta}_0^{TT}\boldsymbol{VV}_0^{-1}\boldsymbol{\beta\beta}_0 + \frac{1}{\sigma\sigma^2}\boldsymbol{yy}^{TT}\boldsymbol{yy} - \frac{2}{\sigma\sigma^2}(\boldsymbol{XX}^{TT}\boldsymbol{yy})^{TT}\boldsymbol{\beta\beta} + \frac{1}{\sigma\sigma^2}(\boldsymbol{XX\beta\beta})^{TT}\boldsymbol{XX\beta\beta}$$

Rearranging terms yields

$$\boldsymbol{\beta\beta}^{TT}\left(\boldsymbol{VV}_0^{-1} + \frac{1}{\sigma\sigma^2}\boldsymbol{XX}^{TT}\boldsymbol{XX}\right)\boldsymbol{\beta\beta} - 2\left(\boldsymbol{VV}_0^{-1}\boldsymbol{\beta\beta}_0 + \frac{1}{\sigma\sigma^2}\boldsymbol{XX}^{TT}\boldsymbol{yy}\right)^{TT}\boldsymbol{\beta\beta} + \boldsymbol{\beta\beta}_0^{TT}\boldsymbol{VV}_0^{-1}\boldsymbol{\beta\beta}_0 + \frac{1}{\sigma\sigma^2}\boldsymbol{yy}^{TT}\boldsymbol{yy}$$

We ignore the last two terms as they are constants. We then only need to add and substract the constant term $\boldsymbol{\alpha\alpha}^{TT}\boldsymbol{AA}^{-1}\boldsymbol{\alpha\alpha}$ to complete the square.

Appendix 2

See Appendix 1 for the key matrix identity we will use. We are combining two terms and will end with a single term that is quadratic in $\boldsymbol{\beta}$. Thus we want the $\boldsymbol{u}$ term in the matrix identity to represent $\boldsymbol{\beta}$, as in Appendix 1. We set $\boldsymbol{\alpha\alpha} = \boldsymbol{VV}_0^{-1}\boldsymbol{\beta\beta}_0 + \boldsymbol{XX}^{TT}\boldsymbol{yy}$ and $\boldsymbol{AA} = \boldsymbol{VV}_0^{-1} + \boldsymbol{XX}^{TT}\boldsymbol{XX}$

We begin with the expression $(\boldsymbol{\beta\beta} - \boldsymbol{\beta\beta}_0)^{TT}\boldsymbol{VV}^{-1}(\boldsymbol{\beta\beta} - \boldsymbol{\beta\beta}_0) + (\boldsymbol{yy} - \boldsymbol{XX\beta\beta})^{TT}(\boldsymbol{yy} - \boldsymbol{XX\beta\beta})$

Expanding the expression we find

$$\boldsymbol{\beta}\boldsymbol{\beta}^{TT}\boldsymbol{V}\boldsymbol{V}_{0}^{-1}\boldsymbol{\beta}\boldsymbol{\beta} - 2(\boldsymbol{V}\boldsymbol{V}_{0}^{-1}\boldsymbol{\beta}\boldsymbol{\beta}_{0})^{TT}\boldsymbol{\beta}\boldsymbol{\beta} + \boldsymbol{\beta}\boldsymbol{\beta}^{TT}\boldsymbol{V}\boldsymbol{V}_{0\ 0}^{-1}\boldsymbol{\beta}\boldsymbol{\beta}_{0} + \boldsymbol{y}\boldsymbol{y}^{TT}\boldsymbol{y}\boldsymbol{y} - 2(\boldsymbol{X}\boldsymbol{X}^{TT}\boldsymbol{y}\boldsymbol{y})^{TT}\boldsymbol{\beta}\boldsymbol{\beta} + (\boldsymbol{X}\boldsymbol{X}\boldsymbol{\beta}\boldsymbol{\beta})^{TT}\boldsymbol{X}\boldsymbol{X}\boldsymbol{\beta}\boldsymbol{\beta}$$

Rearranging terms yields

$$\boldsymbol{\beta}\boldsymbol{\beta}^{TT}(\boldsymbol{V}\boldsymbol{V}_{0}^{-1} + \boldsymbol{X}\boldsymbol{X}^{TT}\boldsymbol{X}\boldsymbol{X})\boldsymbol{\beta}\boldsymbol{\beta} - 2(\boldsymbol{V}\boldsymbol{V}_{0}^{-1}\boldsymbol{\beta}\boldsymbol{\beta}_{0}+bb^{TT}yy)^{TT}\beta\beta + \beta\beta_{0\ 0}^{TT}VV^{-1}\beta\beta_{0} + yy^{TT}yy$$

The first two terms are $uu^{TT}AAuu - 2\alpha\alpha^{TT}uu$ so we add and subtract $\alpha\alpha^{TT}AA^{-1}\alpha\alpha$ in order to complete the square. After applying the identity we have

$$(\beta\beta - AA^{-1}\alpha\alpha)^{TT}AA(\beta\beta - AA^{-1}\alpha\alpha) + \beta\beta_{0\ 0}^{TT}VV^{-1}\beta\beta_{0} + yy^{TT}yy - \alpha\alpha^{TT}AA^{-1}\alpha\alpha$$

using A and α to keep the expression simple.

Noting that $VV_{NN} = AA^{-1}$ and $\beta\beta_{NN} = AA^{-1}\alpha\alpha$ we have

$$(\beta\beta - \beta\beta_{NN})^{TT}VV_{NN}^{-1}(\beta\beta - \beta\beta_{NN}) + \beta\beta_{0\ 0}^{TT}VV^{-1}\beta\beta_{0} + yy^{TT}yy - \beta\beta_{NN\ NN}^{TT}VV^{-1}\beta\beta_{NN}$$

which is what we sought to prove. Note that we do not ignore the constants in this case as we did in the known variance case. The difference in this case is that we have an inverse gamma parameter in the exponent, so the additional terms are not constants as in the known variance case.

Appendix 3

To see that $pp(\boldsymbol{y}\boldsymbol{y}|\boldsymbol{x}\boldsymbol{x},\mathcal{D}\mathcal{D}) = \mathcal{T}\mathcal{T}$ $\frac{bb_{NN}}{NN,\ aa_{NN}}(1 + {}_{NN}\boldsymbol{X}\boldsymbol{X}^{TT}),\ {}_{NN}\boldsymbol{V}$, note that $\boldsymbol{V}yy|XX\beta\beta$

$$pp(\boldsymbol{y}\boldsymbol{y}|\boldsymbol{x}\boldsymbol{x},\mathcal{D}\mathcal{D}) = \int_{0}^{\infty}\int_{-\infty}^{\infty} pp(\boldsymbol{y}\boldsymbol{y}|\boldsymbol{\beta}\boldsymbol{\beta},\sigma\sigma^{2})pp(\boldsymbol{\beta}\boldsymbol{\beta},\sigma\sigma^{2}|\boldsymbol{y}\boldsymbol{y})dd\boldsymbol{\beta}\boldsymbol{\beta}dd\sigma\sigma^{2}$$

$$= \int_{0}^{\infty}\int_{-\infty}^{\infty} NN(\boldsymbol{X}\boldsymbol{X}^{TT}\boldsymbol{\beta}\boldsymbol{\beta},\sigma\sigma^{2})NN\boldsymbol{\beta}\boldsymbol{\beta}|\boldsymbol{\beta}\boldsymbol{\beta}_{NN},\sigma\sigma^{2}\boldsymbol{V}\boldsymbol{V}_{NN}IIII(\sigma\sigma^{2}|aa_{NN},\ bb_{NN})dd\boldsymbol{\beta}\boldsymbol{\beta}dd\sigma\sigma^{2}$$

$$\int_{0}^{\infty}\int_{-\infty}^{\infty} NN(\boldsymbol{X}\boldsymbol{X}^{TT}\boldsymbol{\beta}\boldsymbol{\beta},\sigma\sigma^{2})NN\boldsymbol{\beta}\boldsymbol{\beta}|\boldsymbol{\beta}\boldsymbol{\beta}_{NN},\sigma\sigma^{2}\boldsymbol{V}\boldsymbol{V}_{NN}dd\boldsymbol{\beta}\boldsymbol{\beta}\ IIII(\sigma\sigma^{2}|aa_{NN},\ bb_{NN})dd\sigma\sigma^{2}$$

We have already shown that for the known variance case that

$$\int_{-\infty}^{\infty} NN(\boldsymbol{X}\boldsymbol{X}^{TT}\boldsymbol{\beta}\boldsymbol{\beta},\ \sigma\sigma^{2})NN\boldsymbol{\beta}\boldsymbol{\beta}|\boldsymbol{\beta}\boldsymbol{\beta}_{NN},\sigma\sigma^{2}\boldsymbol{V}\boldsymbol{V}_{NN}dd\boldsymbol{\beta}\boldsymbol{\beta} = NN\boldsymbol{\beta}\boldsymbol{X}\boldsymbol{X}^{TT}\boldsymbol{\beta}\boldsymbol{\beta}_{NN},\ \mathbb{w}^{2}(1 + \boldsymbol{X}\boldsymbol{X}^{TT}\boldsymbol{V}\boldsymbol{V}_{NN}\boldsymbol{X}\boldsymbol{X})$$

Substituting we have

$$\int_{0}^{\infty} NN\boldsymbol{\beta}\boldsymbol{X}\boldsymbol{X}^{TT}\boldsymbol{\beta}\boldsymbol{\beta}_{NN},\ \mathbb{w}^{2}(1 + \boldsymbol{X}\boldsymbol{X}^{TT}\boldsymbol{V}\boldsymbol{V}_{NN}\boldsymbol{X}\boldsymbol{X})IIII(\sigma\sigma^{2}|aa_{NN},\ bb_{NN})dd\sigma\sigma^{2}$$

To keep the notation simples, let $\boldsymbol{V}\boldsymbol{V}^{*} = 1 + \boldsymbol{X}\boldsymbol{X}^{TT}\boldsymbol{V}\boldsymbol{V}_{NN}\boldsymbol{X}\boldsymbol{X}$ and $\phi\phi = (\boldsymbol{y}\boldsymbol{y} - \boldsymbol{X}\boldsymbol{X}^{TT}\boldsymbol{\beta}\boldsymbol{\beta}_{NN})^{TT}\boldsymbol{V}\boldsymbol{V}^{*-1}(\boldsymbol{y}\boldsymbol{y} - \boldsymbol{X}\boldsymbol{X}^{TT}\boldsymbol{\beta}\boldsymbol{\beta}_{NN})$ Then the above expression is

$$\frac{bb_{NN}^{aa_{NN}}}{(2\pi\pi)^{\frac{DD}{2}}|\boldsymbol{VV}^{*}|^{\frac{1}{2}}\Gamma(aa_{NN})}\int_{0}^{\infty}\left(\frac{1}{\sigma\sigma^{2}}\right)^{aa_{NN}+\frac{DD}{2}+1}\exp\left(-\frac{1}{\sigma\sigma^{2}}\left(bb_{NN}+\frac{\phi\phi}{2}\right)\right)dd\sigma\sigma^{2}$$

Let $\tau\tau=\frac{1}{\sigma\sigma^{2}}$. Then $\tau\tau$ has the same range and $dd\tau\tau=-\frac{2}{\sigma\sigma^{4}}dd\sigma\sigma^{2}$. Substituting, the above expression becomes

$$\frac{bb_{NN}^{aa_{NN}}}{(2\pi\pi)^{\frac{DD}{2}}|\boldsymbol{VV}^{*}|^{\frac{1}{2}}\Gamma(aa_{NN})}\int_{0}^{\infty}(\tau\tau)^{aa_{NN}+\frac{DD}{2}+1}\exp\left(-\tau\tau\left(bb_{NN}+\frac{\phi\phi}{2}\right)\right)\frac{1}{2\tau\tau^{2}}dd\tau\tau$$

$$=\frac{bb_{NN}^{aa_{NN}}}{(2\pi\pi)^{\frac{DD}{2}}|\boldsymbol{VV}^{*}|^{\frac{1}{2}}\Gamma(aa_{NN})}\int_{0}^{\infty}(\tau\tau)^{aa_{NN}+\frac{DD}{2}-1}\exp\left(-\tau\tau\left(bb_{NN}+\frac{\phi\phi}{2}\right)\right)dd\tau\tau$$

The integrand has the form of a gamma distribution, the probability density function for which is defined as

$$pp(xx)=\frac{1}{\Gamma(kk)\theta\theta^{kk}}xx^{kk-1}\exp\left(-\frac{xx}{\theta\theta}\right)$$

Letting $kk=aa_{NN}+\frac{DD}{2}$ and $\theta\theta=\left(bb_{NN}+\frac{\phi\phi}{2}\right)^{-1}$, we can write

$$\frac{bb_{NN}^{aa_{NN}}\Gamma\left(aa_{NN}+\frac{DD}{2}\right)\left(bb_{NN}+\frac{\phi\phi}{2}\right)^{-\left(aa_{NN}+\frac{DD}{2}\right)}}{(2\pi\pi)^{\frac{DD}{2}}|\boldsymbol{VV}^{*}|^{\frac{1}{2}}\Gamma(aa_{NN})}\int_{0}^{\infty}\frac{1}{\Gamma\left(aa_{NN}+\frac{DD}{2}\right)\left(bb_{NN}+\frac{\phi\phi}{2}\right)^{-\left(aa_{NN}+\frac{DD}{2}\right)}}(\tau\tau)^{aa_{NN}+\frac{DD}{2}-1}\exp\left(-\tau\tau\left(bb_{NN}+\frac{\phi\phi}{2}\right)\right)dd\tau\tau$$

The range of a gamma distribution is the nonnegative real numbers, so the integrand evaluates to *1*, leaving

$$\frac{bb_{NN}^{aa_{NN}}\Gamma\left(aa_{NN}+\frac{DD}{2}\right)\left(bb_{NN}+\frac{\phi\phi}{2}\right)^{-\left(aa_{NN}+\frac{DD}{2}\right)}}{(2\pi\pi)^{\frac{DD}{2}}|\boldsymbol{VV}^{*}|^{\frac{1}{2}}\Gamma(aa_{NN})}$$

This expression is proportional to

$$\frac{\Gamma\left(\frac{2aa_{NN}+NN}{2}\right)\left(1+\frac{(yy-\boldsymbol{XX}^{TT}\boldsymbol{\beta\beta}_{NN})^{TT}\left(\frac{bb_{NN}}{aa_{NN}}(1+\boldsymbol{XX}^{TT}\boldsymbol{VV}^{*}\boldsymbol{XX})\right)^{-1}(yy-\boldsymbol{XX}^{TT}\boldsymbol{\beta\beta}_{NN})}{2aa_{NN}}\right)^{-\left(aa_{NN}+\frac{DD}{2}\right)}}{\pi\pi^{\frac{DD}{2}}\left(2aa_{NN}\cdot\frac{bb_{NN}}{aa_{NN}}\right)^{\frac{1}{2}}\Gamma(aa_{NN})}$$

which is a multivariate Student's t distribution with mean $\boldsymbol{XX}^{TT}\boldsymbol{\beta\beta}$, variance $\frac{bb_{NN}}{aa_{NN}}(1+$

$\boldsymbol{XX}^{TT}\boldsymbol{VV}_{NN}\boldsymbol{XX}$), and $2aa_{NN}$ degrees of freedom.

References

Boehm, B., A. Hira, K. Qi, and E. Venson, “Calibrating Use Case Points Using Bayesian Analysis,” presented at the 2018 International Cost Estimating and Analysis Association Annual Conference.

Cochran, W. G., 1934,"The distribution of quadratic forms in a normal system, with applications to the analysis of covariance," Mathematical Proceedings of the Cambridge Philosophical Society. 30 (2).

Congdon, P., 2006, *Bayesian Statistical Modelling*, 2nd Edition, Wiley, West Sussex.

Druker, E.R., R.L. Coleman, and P.J. Braxton, “Don’t Let the Financial Crisis Happen to You: Why Estimates Using Power CERs are Likely to Experience Cost Growth,” presented at the 2009 ISPA-SCEA Annual Conference.

Foussier, P.M.M., “The Benefits of the Bayesian Approach vs. the Frequentist Approach when Dealing with Low Data Sample,” presented at the 2010 ISPA-SCEA conference.

Hoffman, P., The Man Who Loved Only Numbers: The Story of Paul Erdos and the Search for Mathematical Truth, Hyperion, 1998, New York, New York.

Smart, C.B., “Enhancing Risk Calibration Methods,” presented at the 2018 International Cost Estimating and Analysis Association Annual Conference.

Smart, C.B., “Covered with Oil: Incorporating Realism in Cost Risk Analysis,” Journal of Cost Analysis and Parametrics, 2015.

Smart, C.B., 2014, “Bayesian Parametrics: How to Develop a CER with Limited Data and Even Without Data,” presented at the 2014 International Cost Estimating and Analysis Association Annual Conference.

Taleb, N.N., *Skin in the Game*: *Hidden Asymmetries in Daily Life*, Random House, New York, 2018.