SAS® GLOBAL FORUM 2020 #SASGF

Paper 4311-2020

Incorporating Auxiliary Information into Your Model Using Bayesian Methods in SAS® Econometrics

Matthew Simpson, SAS Institute Inc.

ABSTRACT

In addition to data, analysts often have available useful auxiliary information about inputs into their model—for example, knowledge that high prices typically decrease demand or that sunny weather increases outdoor mall foot traffic. If used and incorporated correctly into the analysis, the auxiliary information can significantly improve the quality of the analysis. But this information is often ignored. Bayesian analysis provides a principled means of incorporating this information into the model through the prior distribution, but it does not provide a road map for translating auxiliary information into a useful prior. This paper reviews the basics of Bayesian analysis and provides a framework for turning auxiliary information into prior distributions for parameters in your model by using SAS® Econometrics software. It discusses common pitfalls and gives several examples of how to use the framework.

INTRODUCTION

Modern statistical analysis excels at generating insights from data, but these tools can take into account only the inputs that you provide them with. Often you have important information about the problem in the form of vague intuitions, but not in a quantifiable form that you can plug directly into your model. Further, even when you do have concrete auxiliary data relevant to your problem, it might not be straightforward to combine those data with your original data in a larger model. Bayesian analysis is an incredibly powerful means of taking into account various forms of auxiliary information through the so-called prior distribution. However, it is not straightforward to construct this prior, and a poorly constructed prior can yield significantly worse inferences than the ones you make by disregarding the auxiliary information altogether.

This paper provides a conceptual framework for thinking about the prior in order to make it easier for you to construct custom priors by using various sources of auxiliary information. The key is to transform the model and data in a variety of ways in order to make it easier to think about the model parameters and what they imply about observables. The general workflow is as follows: 1) convert the model and data to something that is easy for the analyst to have intuitions about; 2) convert those intuitions to numbers; and 3) convert those numbers to a prior distribution on the transformed version of the problem. To facilitate this process, the paper presents a number of rules of thumb for transforming the model and data into objects that are easier to think about and for converting intuitions into prior distributions.

The rest of the paper is organized as follows. The section "The Bayesian Story and its Discontents" sketches the subjective Bayesian philosophy of statistics and the problems with naively applying it, especially problems associated with the prior distribution. Then the section "Rules to Derive By" discusses some rules of thumb for overcoming these problems and translating various sorts of auxiliary information into prior distributions. Next "Example 1" introduces an example in order to illustrate how to apply the rules in a regression model. "Example 2" then introduces a new data source to illustrate how to apply the rules in probit regression. Continuing with the data in the original example, "Example 3" shows how to use the rules to construct an informative prior by incorporating the results of the probit regression, this time in the context of a count regression. Finally, the "Discussion" section summarizes and reviews the ideas in the paper.

THE BAYESIAN STORY AND ITS DISCONTENTS

The subjective Bayesian philosophy of statistics starts by identifying epistemic uncertainty with probability. That is, your uncertainty about a set of propositions should follow the rules of probability. Then Bayes' rule provides a convenient way to revise beliefs in light of new information. In its simplest form, Bayes' rule is

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

In practice B is the observed data, denoted by $\mathcal{D}$, and A is the model parameters, denoted by $\boldsymbol{\theta}$. This leads to the usual form of Bayes' rule:

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}$$

In this notation, $p(\cdot)$ is the probability density of the enclosed quantity, so $p(\mathcal{D}|\boldsymbol{\theta})$ is the probability density of the data given any model parameters (that is, the likelihood), and $p(\boldsymbol{\theta})$ is the probability density of just the model parameters (that is, the prior density). To a Bayesian analyst, the likelihood tells you how you should update your beliefs about the data that you expect to see once you observe the model parameter, whereas the prior represents your beliefs about the model parameter before you see any data. Combined, they tell you how you should update your beliefs about future observations conditional on the observations that you have already seen. Then the posterior density, $p(\boldsymbol{\theta}|\mathcal{D})$, represents your beliefs about the model parameters after seeing the data. As in any statistical analysis, any questions that the analyst wants to answer need to be converted to statements about the model parameters.

This approach yields two major benefits. First, in principle, it enables the analyst to take into account other sources of information that are not normally considered as part of the "data," including relevant prior beliefs. This should yield more reliable inferences, in theory. Second, it enables the analyst to rigorously make probability statements. For example, in a forecast, the analyst might be able to rigorously say, "The probability that we meet revenue targets in the fourth quarter, conditional particular data sources, a particular model, and a particular prior is about 0.64." Such posterior probabilities are usually more directly relevant to decisions than, for example, the result of a hypothesis test.

Criticisms of Bayesian inference often focus on the prior distribution. Where does it come from? In practice, it can seem completely made up. This criticism is not completely unfair, but Bayesians have a glib response: it comes from the same place as the likelihood. They're both made up, in the sense that the analyst chooses them on the basis of some combination of judgment, experience, and convenience rather than, for example, basing the choice on some set of undeniable axioms. Even formal model selection criteria require judgment to use—for example, which criterion should you use? But according to Bayesians, you *always* have prior information. How often do you check to see whether your regression estimates are "reasonable"? The problem is that it is not obvious how to translate this information into a mathematically precise prior distribution. The Bayesian story is silent on this question, because it assumes that the prior already exists. Similarly, classical inference is silent on how to choose your significance level, because the theories of hypothesis testing and confidence intervals assume that it already exists. Answers to how, precisely, to deal with these issues help define the various flavors of Bayesian statistics, and indeed, the various non-Bayesian approaches to statistical inference.

The glib Bayesian response is insightful once it is fully fleshed out, but it does not completely justify the use of Bayesian methods. Even if you believe the arguments that subjective uncertainty follows the rules of probability, this does not mean that your statistical methods must be Bayesian. It might not be worth the hassle to translate all your auxiliary information into a mathematically precise prior distribution, and because of errors introduced in this process, a non-Bayesian method might actually yield inferences closer to the "true Bayesian" inferences.

But taking into account more information is appealing, as is the ability to make probability statements about quantities of interest. Even if you do not believe the Bayesian story, you can reconceptualize the benefits of the prior distribution in terms of regularization, or the bias-variance trade-off. To make this work, it is key to figure out how to operationalize the prior—how to construct and use it in real-world situations. Constructing a prior is no different from constructing the likelihood in a lot of ways. It typically is not obvious what the "correct" likelihood is, but often you can settle on a "good enough" likelihood while acknowledging its limitations for many uses.

RULES TO DERIVE BY

The previous section is careful to mention that the probability statements that are the output of Bayesian analysis are *conditional* on a variety of things, not just the observed data. It is obvious, but also worth making explicit, that posterior probabilities are also conditional on the likelihood and the prior—if you change either, then the posterior probabilities change. Together, the likelihood and the prior form a *model of your uncertainty* about the problem. From the subjective Bayesian perspective, the model is both pieces simultaneously. They are in some sense inseparable, and in fact the prior often does not make much sense outside the context of the likelihood. But it is usually easier to understand and construct the likelihood portion of the model than the prior, so it is OK to start there. The remainder of this paper assumes that a suitable likelihood has already been chosen, but keep in mind that this is not necessarily always true, and indeed the distinction between prior and likelihood is ambiguous in many contexts. All this leads to the first rule of thumb that you can use to construct prior densities.

Rule 1 *To choose a prior for a parameter, first understand how the parameter affects the distribution of observables.*

What does the regression coefficient imply about the observed responses, compared to the regression coefficient for other covariates? What about the error variance? These questions are not always easy to answer in any precise form. But there are several tricks that you can use to build your intuition. The key in each case is to start with something you have strong intuitions or relevant auxiliary information about—typically the distribution of the data. Then you convert that information to relevant information about the parameters in question.

The most general trick is transformation. Sometimes the scale of the data or the parameters is not easy to think about, so you should transform them to a quantity that is easier to think about.

Rule 2 *Transform quantities in the model to make them easier to think about.*

In practice this rule takes many forms, depending on the quantities in question. A simple but powerful example is to focus on standard deviations and correlations instead of variances and covariances. Standard deviations are already in the same units as the data, and the 68/95/99 rule for normal distributions allows for easy interpretation. If a distribution is approximately normal, then approximately 68% of the distribution is within one standard deviation of the mean, approximately 95% is within two standard deviations, and approximately 99% is within three standard deviations. Similarly, correlations are easier to understand than covariances because they are unitless. A correlation of 0.5 means the same thing no matter the units of the variables, so you can safely abstract away from units when constructing a prior.

Rule 3 *Construct priors for standard deviations and correlations instead of variances and covariances.*

A covariance matrix larger than 2×2 should be handled differently because of the positive definiteness constraint. Typically, a good strategy is to write the prior in terms of standard deviations and the correlation matrix, but these details are beyond the scope of this paper.

It is also useful to transform the data, most often by centering and scaling; that is, subtract the mean of the variable and then divide by its standard deviation to standardize the variable.

Rule 4 *Center and scale continuous covariates, and when applicable the response, in order to make regression coefficients easier to interpret.*

This rule ensures that each regression coefficient can be interpreted on a similar scale: “A change of one standard deviation in this covariate should predict how much of a change in the response?” Technically this is “using the data twice,” since you are using it in both the prior and the likelihood. This makes many Bayesians uncomfortable, because the prior is supposed to be completely independent from the data that you are analyzing. Preventing yourself from using the data to select the prior will make you look more like a Bayesian—as if you are following the procedure of Bayesian inference. But in practice, using the data to help operationalize the prior can help you get closer to the correct Bayesian answer (conditional on the likelihood, your prior information, and so on). So although it is not ideal, standardizing covariates is so useful in constructing priors that it is typically worth it. In principle, you can center and scale with population values if available, such as from census data, to obtain the same benefits for prior construction without using the data twice. The main caveat here is that if your sample is not representative of the population you

are trying to make inferences about, then centering and scaling by the sample means and standard deviations might do more harm than good—for example, if your sample is too small.

Centering and scaling does not make much sense for categorical variables. So instead of transforming them, it is helpful to construct a base case. For example, assume that an observation is in a particular set of categories and the continuous covariates are set to particular values. What do you expect the observables to look like? How do you expect them to change if you move the observation to a different category? This can be complicated, but an easier process is to assume that the mean for the base case is some intuitive value. The examples in this paper illustrate this.

Rule 5 *Do not transform categorical covariates. Instead, construct a base case and think about what you expect for that base case and for changes from the base case to different categories.*

The basic idea of a base case works for continuous variables as well, especially after centering and scaling. For example, in nonlinear models or models with interactions, the impact of a covariate can depend on the values of the other covariates.

Rule 6 *If the impact of a continuous covariate depends on the values of other covariates, think about the impact of that covariate in the context of an intuitive base case.*

The key to choosing a base case is to pick one that is easy to think about, though this often depends on the model.

Another case where transforming is often not helpful is log-log models. That is, if the response is the log of the response that you care about, and the covariate is the log of the covariate that you care about, then the regression coefficient has an easy interpretation as an elasticity: "A 1% change in the covariate predicts a β% change in the response."

Rule 7 *In log-log models, do not center and scale the response or logged covariates. Instead, interpret regression coefficients as elasticities.*

This rule is particularly useful for economic variables. Often there are published papers that estimate elasticities directly relevant to your problem, and you can use them to help inform your prior.

When you do all these transformations to help you think about the model, you ultimately have to choose priors for the parameters of the transformed model. These priors depend to a large extent on the model, but there are some concrete choices that apply fairly generally. First, before trying to incorporate your background knowledge, you should try to construct default priors.

Rule 8 *Construct default, weakly informative priors first, even if you intend to do the analysis with informative priors.*

It is often useful to see how much prior or auxiliary information is driving your inference. This is not necessarily a bad thing, but it is worth knowing what drives your inferences. Also, recall that from the Bayesian perspective, the prior is part of a model of your uncertainty. In general, it is good practice to compare your favored model against reasonable defaults, and fitting the model with weakly informative priors is one way to do this for your uncertainty model.

It can be attractive to assume that a "flat" prior is a good default. It seems uninformative, because a uniform distribution implies that each area of the parameter space is equally likely, a priori. This turns out to be a bad idea for two main reasons. The smaller issue is that it can often cause computational problems, especially for unbounded parameter spaces and for complicated models. The larger issue is that the intuition that "flat" equals "uninformative" is just wrong. It turns out that *no* prior is globally uninformative. Every prior is informative for some questions. The classic way to see this is transformation. A flat prior for θ implies a prior on θ^2 that puts a lot of mass near zero. A prior can be more or less informative for a particular question, however. The next rule of thumb summarizes this.

Rule 9 *No prior is globally uninformative. Instead, default priors should be weakly informative for the questions that the analyst is trying to answer.*

Without looking at a particular model, you can operationalize default priors to some extent. Generally, they should be centered on the values that you would expect if you had a "default" state of knowledge, and they should be spread out very far. In practice, this depends on the type of parameter that you are considering. Regression coefficients are easiest, as the next rule shows.

Rule 10 *In the absence of other information, a good default prior for a regression coefficient is $\beta_j \sim \mathrm{N}(m, s^2)$, where m is what you expect the estimated coefficient to be, and s is chosen so that you do not expect β_j to be more extreme than $m \pm s$, a priori.*

According to the 68/95/99 rule, this prior says that β_j is more extreme than you thought was possible about 32% of the time. This illustrates just how a weakly informative prior is intended to be—it lightly discourages crazy values of the parameters. This prior assumes that the regression parameter is of direct interest. If you are interested in some function of the regression parameters, use your auxiliary information about that quantity to construct your default priors.

It might be tempting to strictly bound the prior between two extreme values, but in general this is not a good idea. It can cause computational issues, but more important, it is better to allow the model to go into extreme regions of the parameter space if the data are strongly telling it to. For example, you might think the coefficient for the price covariate in a sales regression should be negative, but what if the good is a Giffen good, which consumers buy more of as the price rises? Or if consumers use price as a signal of quality or to signal how wealthy they are? The main exception is for parameters that must be defined in a constrained space by their definition, such as standard deviations and correlations.

Rule 11 *Don't constrain parameters in the prior. Instead, construct priors that regularize away from values that seem impossible but still allow them.*

You can also choose some default values for m and s in Rule 10, though this depends on the model and the type of covariate. The examples cover this detail.

Another common type of parameter is a scale parameter, such as variances, standard deviations, and so on. The conventional wisdom is to use an inverse gamma prior on the variance. However, the inverse gamma prior can be highly informative in ways that are typically undesirable, making it a poor choice for a default prior (see, for example, Gelman 2006). A better choice is a positive truncated normal distribution on the standard deviation, which allows for posterior standard deviations to be arbitrarily close to zero but still penalizes values that seem far too large relative to prior expectations.

Rule 12 *In the absence of other information, use an $\mathrm{N}^+(0, s^2)$ prior on standard deviations, with s set to the upper bound of what you reasonably expect a priori. For other scale parameters, transform to a scale similar to the standard deviation first, then use the $\mathrm{N}^+(0, s^2)$ prior for the transformed parameter.*

By the 68/95/99 rule, this prior implies that the standard deviation is larger than the largest value you could reasonably expect about 32% of the time, which again is only weakly informative about the likely value of that parameter. If in your particular application you do not expect standard deviation values near zero, you can add a mean parameter to the positive truncated normal prior to center it on larger values, such as $\mathrm{N}^+(m, s^2)$, where $m > 0$.

Normal tails are very light, and this affects how the model deals with outliers. When the prior has light tails, the posterior takes outlier observations into account very seriously so that one very large observation can strongly influence the posterior (see, for example, O'Hagan 1979). Typically, it is more desirable to achieve "robust" behavior—that is, when an observation is very extreme relative to the prior and other observations, the posterior places less weight on it. In practice, you can achieve this behavior by using fatter-tailed priors, such as by replacing normals with T distributions.

Rule 13 *To make inferences more robust to outliers, use fatter-tailed priors. For example, replace $\mathrm{N}(m, s^2)$ priors with $\mathrm{T}_d(m, s^2)$ priors and set d to a value somewhere from 3 to 7.*

The degrees of freedom parameter, d, controls how fat the tails are. A larger value implies thinner tails and less robust behavior. Setting d to too small a value can cause computational problems and sometimes break the assumptions that are required for doing Markov chain Monte Carlo simulation (see, for example, Ghosh, Li, and Mitra 2018). So in the absence of a compelling reason to do something different, you should generally set d to at least 3.

EXAMPLE 1: LOG SALES REGRESSION WITH DEFAULT PRIORS

To illustrate how to use the rules of thumb in the previous section, consider a hypothetical network of 100 car dealerships. This network is considering expanding to one of several new locations, and it wants to forecast the yearly sales of a particular model of four-wheel-drive pickup truck—in particular, what impact the price of the truck will have on those sales. As a baseline, the network wants to fit a regression model that uses last year's sales and price data from the existing dealerships in the network to predict the number of trucks that it will sell, controlling for climate and demographic variables for the region. The available variables for each dealership are the type of region it is (**area_type**: rural, suburban, or urban); the number of people in the region who are at least 18 years old and have at least a bachelor's degree (**pop_bachelors**); the number of people in the region who are at least 18 years old and have less than a bachelor's degree (**pop_below_bachelors**); median household income in the region in dollars (**median_income**); cost of living for the region, as measured by an available index (**cost_of_living**, a positive number); average high temperature in degrees Fahrenheit in the summer months—June, July, and August (**mean_summer_temp**); average high temperature in degrees Fahrenheit in the winter months—December, January, and February (**mean_winter_temp**); average yearly precipitation in inches (**mean_precip**); the number of trucks sold (**sales**); and the posted price in dollars (**price**). The following code generates the hypothetical data set:

```
data trucksales;
   call streaminit(768234);
   rural_intercept = 10;
   urban_intercept = 8;
   suburban_intercept = 9;
   do i = 1 to 100;
      population = rand('POISSON', 50000);
      prop_bachelors = rand('BETA', 10, 30);
      pop_bachelors = INT(prop_bachelors * population);
      pop_below_bachelors = INT((1 - prop_bachelors) * population);
      median_income = INT(exp(log(40000) + .3*rand('NORMAL', 0, 1)));
      price = ROUND(25000 + 1000 * rand('NORMAL', 0, 1), 100);
      cost_of_living = INT(130 + 20*rand('NORMAL', 0, 1));
      mean_summer_temp = INT(85 + 5*rand('NORMAL', 0, 1));
      mean_winter_temp = INT(35 + 8*rand('NORMAL', 0, 1));
      mean_precip = INT(exp(log(22) + .4*rand('NORMAL', 0, 1)));
      rural_idx = rand('NORMAL', 0, 1);
      if rural_idx < -0.7 then area_type = 'rural';
      if rural_idx >  0.7 then area_type = 'urban';
      if abs(rural_idx) <= 0.7 then area_type = 'sub';
      if area_type = 'rural' then intercept = rural_intercept;
      if area_type = 'urban' then intercept = urban_intercept;
      if area_type = 'sub'   then intercept = suburban_intercept;
      xbeta = intercept - 1 + 0.03 * log(pop_bachelors) +
         0.04 * log(pop_below_bachelors) + 0.04 * log(median_income) +
         - 0.5 * log(price) - 0.02 * log(cost_of_living) +
         - 0.02 * log(mean_summer_temp) + 0.3 * log(mean_winter_temp) +
         0.02 * log(mean_precip);
      sales = CEIL(exp(xbeta + 0.05 * rand('NORMAL', 0, 1)));
      output;
   end;
   keep pop_bachelors pop_below_bachelors median_income price cost_of_living
      mean_summer_temp mean_winter_temp mean_precip area_type sales;
run; quit;
```

Technically, there is an endogeneity problem here because it is not clear whether the price differences are from different demand curves, different supply curves, or both. For the purposes of this and the later examples, ignore this issue. For example, suppose that the supply curve is perfectly elastic and constant across dealerships, since they are all part of the same network. Keep in mind that these assumptions are very strong and that a more sophisticated model would be required if they do not hold. The following code generates a summary of the data set, which is shown in Figure 1:

```
proc summary data = trucksales print maxdec=2;
   var pop_bachelors pop_below_bachelors median_income cost_of_living
      mean_summer_temp mean_winter_temp mean_precip price sales;
run; quit;

proc summary data = trucksales print;
   class area_type;
run; quit;
```

Figure 1 Summary of Climate, Demographic, and Dealership Data

The SUMMARY Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
pop_bachelors	100	12118.38	2956.35	6684.00	20223.00
pop_below_bachelors	100	37857.67	2969.18	29703.00	43258.00
median_income	100	44012.19	13115.21	18261.00	80122.00
cost_of_living	100	127.36	20.26	78.00	176.00
mean_summer_temp	100	84.70	5.16	71.00	95.00
mean_winter_temp	100	34.19	8.18	11.00	60.00
mean_precip	100	23.83	12.21	5.00	92.00
price	100	25020.00	952.72	22600.00	27500.00
sales	100	177.60	123.79	48.00	469.00

The SUMMARY Procedure

area_type	N Obs
rural	22
sub	52
urban	26

When you fit the regression model, you have several choices to make. Many of the variables are positive constrained—the population variables, income, cost of living, precipitation, price, and sales. It is reasonable to log-transform these variables, especially to take advantage of thinking about them as elasticities in order to construct priors (that is, Rule 7). The temperature variables are not positive constrained, so instead Rule 4 suggests that you should center and scale them. The next bit of code performs all these transformations, then generates a summary of the transformed data set shown in Figure 2:

```
data trucksales_log;
   set trucksales;
   log_pop_bachelors = log(pop_bachelors);
   log_pop_below_bachelors = log(pop_below_bachelors);
   log_median_income = log(median_income);
   log_price = log(price);
   log_cost_of_living = log(cost_of_living);
   log_mean_precip = log(mean_precip);
   log_sales = log(sales);
   mean_summer_temp_cs = mean_summer_temp;
   mean_winter_temp_cs = mean_winter_temp;
   keep mean_summer_temp_cs mean_winter_temp_cs area_type
      log_pop_bachelors log_pop_below_bachelors log_median_income log_price
      log_cost_of_living log_mean_precip log_sales;
run; quit;

proc standard data = trucksales_log mean=0 std=1 out=trucksales_transformed;
   var mean_summer_temp_cs mean_winter_temp_cs;
run; quit;
```

```
proc summary data = trucksales_transformed print maxdec=2;
   var log_pop_bachelors log_pop_below_bachelors log_median_income log_cost_of_living
      mean_summer_temp_cs mean_winter_temp_cs log_mean_precip log_price log_sales;
run; quit;
```

Figure 2 Summary of Transformed Climate, Demographic, and Dealership Data

The SUMMARY Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
log_pop_bachelors	100	9.37	0.25	8.81	9.91
log_pop_below_bachelors	100	10.54	0.08	10.30	10.67
log_median_income	100	10.65	0.30	9.81	11.29
log_cost_of_living	100	4.83	0.16	4.36	5.17
mean_summer_temp_cs	100	-0.00	1.00	-2.66	2.00
mean_winter_temp_cs	100	0.00	1.00	-2.83	3.15
log_mean_precip	100	3.08	0.43	1.61	4.52
log_price	100	10.13	0.04	10.03	10.22
log_sales	100	4.95	0.69	3.87	6.15

Next, you can fit the regression model by using classical methods just to get a baseline. The rationale for doing this is the same as the rationale for Rule 8: if there is a major difference in the analyses, that is worth knowing even if you do not think that it is necessarily a problem. The following code obtains the classical estimates by using PROC QLIM; they are reported in Figure 3.

```
proc qlim data = trucksales_transformed plots = none;
   class area_type;
   model log_sales = area_type log_pop_bachelors log_pop_below_bachelors
                     log_median_income log_price log_cost_of_living
                     log_mean_precip mean_summer_temp_cs mean_winter_temp_cs;
run; quit;
```

Figure 3 Classical Fit to Transformed Truck Sales Data

The QLIM Procedure

Parameter Estimates

Parameter		DF	Estimate	Standard Error	t Value	Approx Pr > |t|
Intercept		1	8.876976	4.165329	2.13	0.0331
area_type	rural	1	2.006472	0.014212	141.18	<.0001
area_type	sub	1	0.998033	0.011482	86.92	<.0001
area_type	urban	0	0	.	.	.
log_pop_bachelors		1	0.094135	0.097595	0.96	0.3348
log_pop_below_bachelors		1	0.112005	0.296659	0.38	0.7058
log_median_income		1	0.016668	0.016753	0.99	0.3198
log_price		1	-0.676310	0.125230	-5.40	<.0001
log_cost_of_living		1	-0.070649	0.029772	-2.37	0.0176
log_mean_precip		1	0.019825	0.011464	1.73	0.0837
mean_summer_temp_cs		1	0.000821	0.004924	0.17	0.8676
mean_winter_temp_cs		1	0.078446	0.004981	15.75	<.0001
_Sigma		1	0.046566	0.003311	14.06	<.0001

Now you can move on to constructing default priors. At this stage, the priors that you select should be fairly generic and independent of what the variable means, unless you have a very strong reason to make a different choice. So use Rule 10 for regression coefficients, and set $m = 0$. The only reasonable exception in this example is the price coefficient, because a positive price elasticity of demand would be very surprising. However, as Rule 8 says, it is a good idea to fit the model with the "standard" default prior anyway, instead of a prior that incorporates this extra information about the coefficient.

For all the logged covariates, you can use Rule 7 and construct a generic prior for all elasticities. Because an elasticity can be positive or negative, m = 0. For s, first consider what would be a very surprising large (or small) value for an elasticity. In many cases, estimated demand elasticities are between –1 and 1—for example, a 1% change in the covariate typically results in less than a 1% change in the amount of the product sold—though sometimes elasticities are more extreme than that, depending on the product and covariate. An elasticity of ± 4 would be very surprising in most contexts, so call that the most extreme value you expect. This prior is weakly informative, since it expects to see an elasticity larger in magnitude than the most extreme value you expect about 32% of the time.

Next, consider the centered and scaled covariates. The easiest way to construct a default prior for these coefficients is to think in terms of standard deviations of the response. This gives you a general, unit-free perspective to work with. So if, for example, the mean summer high temperature increased by one standard deviation, how many standard deviations do you expect log sales to change by? Again, setting m = 0 makes sense—it could increase or decrease sales, but by default you do not know. A ± 4 standard deviation change in log sales would be very surprising, so again you can use that as your choice for s. From Figure 2, you can see that the standard deviation is about 0.69, so set $s = 4 \times 0.69 \approx 2.76$.

The next parameters to consider are the dummy variables that are associated with the **area_type** variable. For a default prior, you should have no reason to distinguish among the groups, so each coefficient should be centered on m = 0. To choose s, a good default choice is to suppose that group membership has about the same impact on the response as a change of one standard deviation in a continuous covariate. This implies that s = 2.76.

Finally, you need to choose a prior for the intercept, which under the default dummy encoding in the QLIM procedure corresponds to the intercept for the urban area type. This is trickier, because the value of the intercept varies widely depending on the values of the slopes. So Rule 1 requires you to apply Rules 5 and 6. If you standardized *all* the covariates in the model, then the intercept would be directly interpretable as the unconditional mean of the response variable. In that case, a reasonable default prior would set m equal to the sample mean of the response—that is, 4.95. Typically, you have a combination of dummy variables, standardized continuous covariates, and nonstandardized continuous covariates, as in this example. This makes interpreting the intercept difficult, so it is usually better to set m to the classical estimate of the intercept—that is, 8.88.

To take into account the additional prior uncertainty in the intercept, the value of s should generally be much larger than the values that you choose for the slope coefficients. The largest value of s for any of the regression coefficients was s = 4, so for the intercept s = 100 adds about two orders of magnitude more prior variation. This prior for the intercept builds in an assumption that you are not directly interested in it—that is, it is a nuisance parameter. If you are directly interested in the intercept parameter, then instead of using the classical estimate for m, you should spend more time thinking clearly about what it means for your problem and what you know about it a priori. This approach can be generalized: you should typically put more effort into constructing priors for the parameters that directly matter for your inferential question.

Finally, you need a prior on the error variance—or by Rule 3, the error standard deviation. Rule 12 suggests a good default prior, and you need only choose s. In any regression model, the error standard deviation is almost always less than the response variable's standard deviation, so setting s = 0.69 implies that a priori, you expect the error standard deviation to be above the response's standard deviation about 32% of the time.

All of our priors are listed in one convenient place, as follows. Note that these distributions are assumed to be independent of one another. In principle, you can put dependence in your prior distribution, but in practice it can be quite difficult to think about that dependence and make your intuitions mathematically precise.

$$\begin{aligned}
\beta_{\text{intercept}} &\sim \mathrm{N}(8.88, 100^2) \\
\beta_{\text{log_pop_bachelors}}, \beta_{\text{log_pop_below_bachelors}}, \beta_{\text{log_median_income}}, & \\
\beta_{\text{log_cost_of_living}}, \beta_{\text{log_mean_precip}}, \beta_{\text{log_price}} &\sim \mathrm{N}(0, 4^2) \\
\beta_{\text{mean_summer_temp_cs}}, \beta_{\text{mean_winter_temp_cs}}, & \\
\beta_{\text{area_type_rural}}, \beta_{\text{area_type_sub}} &\sim \mathrm{N}(0, 2.76^2) \\
\sigma &\sim \mathrm{N}^{+}(0, 0.69^2)
\end{aligned}$$

The following code fits the model in PROC QLIM by using a joint random walk Metropolis sampler. Figure 4 shows the associated output, including various posterior summaries.

```
proc qlim data = trucksales_transformed plots = none;
   class area_type;
   model log_sales = area_type log_pop_bachelors log_pop_below_bachelors
      log_median_income log_price log_cost_of_living
      log_mean_precip mean_summer_temp_cs mean_winter_temp_cs;
   bayes seed = 72834 ntu = 100 mintune = 20 maxtune = 20 nmc = 10000
      statistics = (summary interval prior);
   prior intercept ~ normal(mean = 8.88, var = 10000);
   prior log_pop_bachelors log_pop_below_bachelors log_median_income
      log_cost_of_living log_mean_precip log_price ~ normal(mean = 0, var = 16);
   prior mean_summer_temp_cs mean_winter_temp_cs
      area_type_rural area_type_sub ~ normal(mean = 0, var = 7.62);
   prior _sigma ~ normal(mean = 0, var = 0.48);
run; quit;
```

Figure 4 Bayesian Fit with Default Priors to Truck Sales Data

The QLIM Procedure

Posterior Summaries

Parameter	N	Mean	Standard Deviation	Percentiles 25%	50%	75%
Intercept	10000	9.2469	4.8299	5.9232	9.0512	12.2599
area_type_rural	10000	2.0083	0.0147	1.9979	2.0079	2.0188
area_type_sub	10000	0.9981	0.0127	0.9902	0.9978	1.0062
log_pop_bachelors	10000	0.0865	0.1179	0.00838	0.0975	0.1683
log_pop_below_bachelors	10000	0.0979	0.3387	-0.1114	0.1214	0.3326
log_median_income	10000	0.0194	0.0178	0.00711	0.0194	0.0318
log_price	10000	-0.6945	0.1315	-0.7790	-0.6964	-0.6078
log_cost_of_living	10000	-0.0684	0.0318	-0.0893	-0.0687	-0.0475
log_mean_precip	10000	0.0176	0.0123	0.0105	0.0177	0.0254
mean_summer_temp_cs	10000	0.000351	0.00516	-0.00284	0.000519	0.00384
mean_winter_temp_cs	10000	0.0792	0.00521	0.0754	0.0790	0.0825
_Sigma	10000	0.0495	0.00317	0.0472	0.0494	0.0516

Posterior Intervals

Parameter	Alpha	Equal-Tail Interval		HPD Interval	
Intercept	0.050	0.2670	18.8094	-0.0949	18.3721
area_type_rural	0.050	1.9808	2.0363	1.9808	2.0360
area_type_sub	0.050	0.9720	1.0236	0.9733	1.0239
log_pop_bachelors	0.050	-0.1629	0.2867	-0.1355	0.3065
log_pop_below_bachelors	0.050	-0.6299	0.6927	-0.5930	0.7114
log_median_income	0.050	-0.0130	0.0556	-0.0151	0.0518
log_price	0.050	-0.9639	-0.4380	-0.9490	-0.4269
log_cost_of_living	0.050	-0.1324	-0.00468	-0.1371	-0.0101
log_mean_precip	0.050	-0.00965	0.0415	-0.00652	0.0431
mean_summer_temp_cs	0.050	-0.0106	0.0103	-0.00899	0.0114
mean_winter_temp_cs	0.050	0.0692	0.0904	0.0684	0.0893
_Sigma	0.050	0.0439	0.0560	0.0438	0.0559

Compared to Figure 3, many of the slope estimates in Figure 4 are attenuated toward zero. This small bit of regularization comes from the weakly informative prior and can protect you from making premature inferences, such as from *p*-hacking or the garden of forking paths (for an explanation of the garden, see Gelman and Loken 2014). If this is an explicit goal of the prior, then you should construct it with that purpose in mind. For example, center the priors on the "null hypothesis," and use smaller prior standard deviations to further regularize the parameters.

EXAMPLE 2: PURCHASING DECISION PROBIT REGRESSION WITH DEFAULT PRIORS

To gather more information, the dealership network wants to fit a probit model by using internal data from an advertising campaign at one dealership. This dealership sent out a flier to 10,000 individuals in its region with an ad for the pickup truck whose sales the network is trying to forecast in Example 1. The fliers were randomized to include one of four possible advertised prices—$20,000, $21,000, $22,000, or $23,000—and were repeatedly sent throughout the year. A recipient would be able to buy the truck at this price only by coming to the dealership with the flier in hand. The dealership recorded whether each individual who received the flier bought a truck at the advertised price over the next year.

Each of the recipients was already in the dealership's advertising database, with several demographic variables recorded. The data set includes the following variables: the race of the recipient (**race**—white, black, Asian, or other), the age of the recipient in years (**age**), whether the recipient is male or female (**sex**—male = 1, female = 0), the amount of time it would take the recipient to drive to the dealership in minutes (**drive_time**), whether the recipient had previously purchased a vehicle at the dealership (**prev_purchase**), the price of the truck in the advertisement in thousands of dollars (**price**), and whether the recipient purchased a truck at the advertised price (**purchase**). The following code generates this data set and summarizes it. The summary is shown in Figure 5.

```
data truck_ad;
   call streaminit(92342);
   do i = 1 to 10000;
      race_idx = rand('NORMAL', 0, 1);
      if race_idx >= 0 then
         do;
            race = 'white';
            age = rand('POISSON', 55);
            intercept = 1.1;
         end;
      if race_idx < 0 then
         do;
            race = 'black';
            age = rand('POISSON', 50);
            intercept = 0.7;
         end;
      if race_idx < -1 then
         do;
            race = 'asian';
            age = rand('POISSON', 60);
            intercept = -2.9;
         end;
      if race_idx < -2 then
         do;
            race = 'other';
            age = rand('POISSON', 55);
            intercept = -1.8;
         end;
      price_idx = rand('uniform', 0, 4);
      price = 20;
      if price_idx > 1 then price = 21;
      if price_idx > 2 then price = 22;
      if price_idx > 3 then price = 23;
      prev_purchase = rand('BINOMIAL', 0.3, 1);
      sex = rand('BINOMIAL', 0.7, 1);
      drive_time = INT(exp(log(60) + 0.5*rand('NORMAL', 0, 1)));
      mu = 7.5 + intercept + 0.1 * sex + 0.001 * age +
         0.002 * prev_purchase - 0.002 * drive_time - 0.5 * price;
      prob = 1 / (1 + exp(-mu));
      purchase = rand('BINOMIAL', prob, 1);
      output;
   end;
   keep race sex age price prev_purchase drive_time purchase;
run; quit;
```

```
proc summary data = truck_ad print maxdec=2;
   var age sex price drive_time prev_purchase purchase;
run; quit;

proc summary data = truck_ad print;
   class race;
run; quit;
```

Figure 5 Summary of Advertisement Data

The SUMMARY Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
age	10000	54.07	8.12	25.00	88.00
sex	10000	0.69	0.46	0.00	1.00
price	10000	21.49	1.12	20.00	23.00
drive_time	10000	67.48	36.64	7.00	371.00
prev_purchase	10000	0.30	0.46	0.00	1.00
purchase	10000	0.08	0.28	0.00	1.00

The SUMMARY Procedure

race	N Obs
asian	1422
black	3369
other	251
white	4958

Your task is to fit a probit model by using these variables to estimate the impact of the advertised price on whether or not a recipient will respond to the ad. This information is then used to inform the prior on the log price variable from Example 1. The probit model assumes that

$$P(y_i = 1) = \Phi(\mathbf{x}_i'\boldsymbol{\beta})$$

for $i = 1, 2, \ldots, 10,000$, where y_i is the recipient's **purchase** variable, $\mathbf{x}_i$ is a vector of the recipient's covariates listed earlier, $\boldsymbol{\beta}$ is a corresponding vector of regression coefficients, and $\Phi()$ is the standard normal cumulative distribution function. Your task for this example is to come up with a reasonable default prior for $\boldsymbol{\beta}$.

First, Rules 4 and 7 apply. Although this model is not a log-log model, the results of the model will be used to inform the prior on an elasticity, so it is convenient to make the mean structures of the two models similar. Another reason not to standardize **price** is that the variable's variation in the data set is in some sense artificial: it was chosen deliberately by the designers of the ad campaign and does not necessarily represent real-world variation in price. The other continuous variables—**age** and **drive_time**—are also positive constrained, but in this model they are easier to think about when centered and scaled than when log-transformed. The following code transforms each variable as appropriate and produces the summary of the transformed data set shown in Figure 6:

```
data truck_ad_cs;
   set truck_ad;
   age_cs = age;
   drive_time_cs = drive_time;
   log_price = log(price);
   keep race sex age_cs drive_time_cs prev_purchase log_price purchase;
run; quit;

proc standard data = truck_ad_cs mean=0 std=1 out=truck_ad_transformed;
   var age_cs drive_time_cs;
run; quit;

proc summary data = truck_ad_transformed print maxdec=2;
   var age_cs sex log_price drive_time_cs prev_purchase purchase;
run; quit;
```

Figure 6 Summary of Advertisement Data

The SUMMARY Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
age_cs	10000	-0.00	1.00	-3.58	4.18
sex	10000	0.69	0.46	0.00	1.00
log_price	10000	3.07	0.05	3.00	3.14
drive_time_cs	10000	0.00	1.00	-1.65	8.28
prev_purchase	10000	0.30	0.46	0.00	1.00
purchase	10000	0.08	0.28	0.00	1.00

Next, the following code fits the model by using classical methods. Figure 7 shows the resulting parameter estimates.

```
proc qlim data = truck_ad_transformed plots = none;
   class purchase race;
   model purchase = race sex age_cs drive_time_cs prev_purchase log_price
      / discrete(dist = normal);
run; quit;
```

Figure 7 Probit Fit to Transformed Advertisement Data

The QLIM Procedure

Parameter Estimates

Parameter		DF	Estimate	Standard Error	t Value	Approx Pr > |t|
Intercept		1	14.742479	1.154745	12.77	<.0001
race	asian	1	-2.045956	0.293640	-6.97	<.0001
race	black	1	-0.171212	0.041751	-4.10	<.0001
race	other	1	-1.239475	0.268596	-4.61	<.0001
race	white	0	0	.	.	.
sex		1	0.054499	0.041671	1.31	0.1909
age_cs		1	0.023252	0.020944	1.11	0.2669
drive_time_cs		1	-0.024131	0.019554	-1.23	0.2172
prev_purchase		1	-0.038599	0.042132	-0.92	0.3596
log_price		1	-5.228477	0.378214	-13.82	<.0001

To construct a prior for the continuous covariates, first think about what they mean in the context of the model. Following Rule 6, suppose that for some base case recipient, $\mathbf{x}_i'\boldsymbol{\beta} = \mu_i$. Now suppose that the age of the recipient increases by one standard deviation. Then the change in the probability that the recipient purchases the truck, denoted by Δ, is given by

$$\Delta = \Phi(\mu_i + \beta_{\text{age\_cs}}) - \Phi(\mu_i)$$

Solving for $\beta_{\text{age\_cs}}$ yields

$$\beta_{\text{age\_cs}} = \Phi^{-1}\left[\Delta + \Phi(\mu_i)\right] - \mu_i$$

where $\Phi^{-1}()$ is the standard normal quantile function. You can use this equation to choose m and s for use in Rule 10. A useful base case is a recipient who is about 50/50 on whether to purchase the advertised truck, which implies $\mu_i = 0$. Now increase this person's age by one sample standard deviation. How much do you expect the probability that the person will purchase the truck to change? Call this value Δ_M. Then, plugging everything into the preceding equation yields a value for m in the prior for $\beta_{\text{age\_cs}}$. A good default is $\Delta_M = 0$, which yields

$$m = \Phi^{-1}(\Delta_M + \Phi(0)) = \Phi^{-1}(0.5) = 0$$

To get a value for s, what is the largest Δ you would expect after a change in **age** of one standard deviation? What about the smallest Δ? Call these Δ_U and Δ_L, respectively. Then identify them with a one-standard-deviation increase

and decrease relative to m, respectively, to get the following equations:

$$m + s = \Phi^{-1}(\Delta_U + 0.5)$$
$$m - s = \Phi^{-1}(\Delta_L + 0.5)$$

Rearranging and plugging in m = 0 yields

$$s = \Phi^{-1}(\Delta_U + 0.5)$$
$$s = -\Phi^{-1}(\Delta_L + 0.5)$$

A good default here is $\Delta_U = 0.3 = -\Delta_L$. Because the maximum possible change in $P(y_i = 1)$ from the base case is ± 0.5, this would be a very large change relative to such a small change in the covariate. Plugging it into the equation yields

$$s = \Phi^{-1}(0.8) \approx 0.84$$
$$s = -\Phi^{-1}(0.2) \approx 0.84$$

Note that $\Phi^{-1}(0.2) = -\Phi^{-1}(0.8)$, so it does not matter whether you take the most positive or most negative extreme value for Δ, but in general you should take the maximum value for s implied by the two equations. All of this reasoning applies for $\beta_{\text{drive\_time}}$ as well, because this is a default prior stated in terms of the sample standard deviation. Thus the independent priors for both variables are

$$\beta_{\text{age\_cs}}, \beta_{\text{drive\_time\_cs}} \sim \text{N}(0, 0.84^2)$$

The elasticity reasoning for **log_price** is complicated in this model, and as discussed earlier, thinking in terms of the sample standard deviation is also not useful. Instead, consider a base case again with **log_price** = $\log 20$, and suppose that the price increases by exactly \$1,000. Then set up the equation for the change in $P(y_i = 1)$:

$$\begin{aligned}\Delta &= \Phi\Big[\mu_i + (\log 21 - \log 20)\beta_{\text{log\_price}}\Big]) - \Phi(\mu_i) \\ &\approx \Phi(\mu_i + 0.05\beta_{\text{log\_price}}) - \Phi(\mu_i)\end{aligned}$$

Again, plug in $\mu_i = 0$ to make the base case as simple as possible, and solve for $\beta_{\text{log\_price}}$ to obtain

$$\beta_{\text{log\_price}} \approx 20\Phi^{-1}(\Delta + 0.5)$$

Now use the same tricks as before. Set $\Delta_M = 0$ to obtain m = 0. Then set $\Delta_U = 0.3 = -\Delta_L$ to obtain

$$\begin{aligned}m + s &= 20\Phi^{-1}(\Delta_U + 0.5) \\ s &= 20\Phi^{-1}(0.8) \approx 20 \times 0.84 \approx 16.8\end{aligned}$$

This produces a default prior of $\beta_{\text{log\_price}} \sim \text{N}(0, 16.8^2)$.

For the coefficients on dummy covariates, **sex**, **prev_purchase**, and the race dummies, most of the work is already done if you can easily compare them to the coefficients that you already have priors for. How much of an impact do you expect a change in the dummy variable to have on $P(y_i = 1)$, relative to a change of one standard deviation in either **age** or **drive_time**? A good default answer is "about the same." This yields the independent priors

$$\beta_{\text{sex}}, \beta_{\text{prev\_purchase}} \sim \text{N}(0, 0.84^2)$$

Finally, you need a prior for the intercept. As in Example 1, a good default for m when all covariates have been standardized is the sample mean response. In this case, that response is a rate and needs to be transformed by the link function—that is, $m = \Phi^{-1}(0.08) = -1.41$. In this example, where some covariates are standardized and others are not, a better default is the classical estimate for the intercept, so m = 14.74. Again, this default prior works best when the intercept is not directly related to your inferential questions.

To choose s, once again set it much larger than the largest slope coefficient standard deviation, which is s = 16.8. For example s = 100 should provide plenty of prior variation to accommodate a wide range of possibilities. This yields the following full set of priors:

$$\beta_{\text{intercept}} \sim \text{N}(14.74, 100^2)$$
$$\beta_{\text{age\_cs}}, \beta_{\text{drive\_time\_cs}}, \beta_{\text{sex}}, \beta_{\text{prev\_purchase}}, \beta_{\text{asian}}, \beta_{\text{black}}, \beta_{\text{other}} \sim \text{N}(0, 0.84^2)$$
$$\beta_{\text{log\_price}} \sim \text{N}(0, 16.8^2)$$

The following code fits the model in PROC QLIM by using the default priors listed earlier and the random walk Metropolis sampler, and it produces the summaries of the posterior shown in Figure 8:

```
proc qlim data = truck_ad_transformed plots = none;
   class purchase race;
   model purchase = race sex age_cs drive_time_cs prev_purchase log_price
      / discrete(dist = normal);
   bayes seed = 2341685 ntu = 100 mintune = 20 maxtune = 20 nmc = 10000
      statistics = (summary interval prior);
   prior intercept ~ normal(mean = 14.74, var = 10000);
   prior age_cs drive_time_cs sex prev_purchase
      race_asian race_black race_other ~ normal(mean = 0, var = 0.71);
   prior log_price ~ normal(mean = 0, var = 283);
run; quit;
```

Figure 8 Bayesian Fit with Default Priors to Advertisement Data

The QLIM Procedure

Posterior Summaries

Parameter	N	Mean	Standard Deviation	Percentiles 25%	50%	75%
Intercept	10000	14.7989	1.0583	14.0680	14.8058	15.5272
race_asian	10000	-1.9033	0.2295	-2.0414	-1.8778	-1.7452
race_black	10000	-0.1728	0.0430	-0.2019	-0.1712	-0.1433
race_other	10000	-1.1579	0.2520	-1.3205	-1.1419	-0.9766
sex	10000	0.0569	0.0421	0.0312	0.0567	0.0862
age_cs	10000	0.0224	0.0214	0.00834	0.0236	0.0369
drive_time_cs	10000	-0.0226	0.0192	-0.0359	-0.0229	-0.00982
prev_purchase	10000	-0.0399	0.0399	-0.0681	-0.0411	-0.0132
log_price	10000	-5.2482	0.3453	-5.4846	-5.2482	-5.0090

Posterior Intervals

Parameter	Alpha	Equal-Tail Interval		HPD Interval	
Intercept	0.050	12.7640	16.9574	12.7596	16.9074
race_asian	0.050	-2.4183	-1.5028	-2.3601	-1.4650
race_black	0.050	-0.2614	-0.0894	-0.2627	-0.0924
race_other	0.050	-1.7089	-0.7257	-1.6214	-0.6623
sex	0.050	-0.0302	0.1375	-0.0285	0.1379
age_cs	0.050	-0.0190	0.0651	-0.0177	0.0653
drive_time_cs	0.050	-0.0593	0.0159	-0.0580	0.0165
prev_purchase	0.050	-0.1188	0.0424	-0.1097	0.0446
log_price	0.050	-5.9548	-4.5839	-5.9167	-4.5661

As in Example 1, this posterior attenuates some of the slope estimates toward zero, though not all of them. But it still serves as a useful baseline for building more informative priors, if that is your goal with this analysis.

EXAMPLE 3: SALES COUNT REGRESSION WITH AN INFORMATIVE PRICE PRIOR

Next, the dealership network wants to improve the original regression model from Example 1 in two ways. First, management wants you to use a count regression model that takes into account the fact that sales is an integer; and second, it wants you to use an informative prior for price that is informed by the probit fit from Example 2.

The basic count regression model is a Poisson regression where

$$y_i \overset{\text{ind}}{\sim} \text{Poisson}(\lambda_i)$$

and $\log \lambda_i = \mathbf{x}_i' \boldsymbol{\beta}$. This model's response variable is **sales** instead of **log_sales**, but the log-link function that defines λ_i enables you to continue to interpret the coefficients for logged covariates as elasticities, at least approximately. This will come in handy when it is time to construct priors. In fact, apart from the coefficient for **log_price**, you can just use the default priors from Example 1 here. This model has no standard deviation or other dispersion parameter, so you can ignore the original regression model's prior for σ.

In an ideal world, a Bayesian analyst might try to build a joint model of the two data sets so that information learned from one data set can spill over into the fit to the other data set, and vice versa. But this is essentially impossible. The two models control for different sets of variables and have different types of response variables, and the advertisement data came from only one of the dealerships represented in the sales data set. Generalizing to every other dealership is nearly impossible, and converting from the context of one set of covariates to another set is even harder—not to mention that such a model can introduce computational difficulties.

Constructing an informative prior for $\beta_{\text{log\_price}}$ from the results of the probit fit is still not straightforward, but not requiring a precise statement of the connection between the two data sets makes it easier. The probit fit tells you *something* about the price elasticity of demand for the truck model, though it is not clear precisely what. You can compensate for this uncertainty by making the prior distribution relatively less certain about the value of the relevant parameter.

The first challenge is to convert information about $\beta_{\text{log\_price}}$ from the probit model to information about an *elasticity*. Suppose that there are N individuals in a given region, and suppose that according to the probit model the ith individual's probability of purchasing the truck is $\Phi(\mu_i)$. Then the expected number of purchases (that is, the expected demand) is $E = \sum_{i=1}^{N} \Phi(\mu_i)$.

To get an elasticity with respect to price, you need the derivative of $\log E$ with respect to **log_price**,

$$\frac{\partial \log E}{\partial \text{log\_price}} = \frac{\sum_{i=1}^{N} \phi(\mu_i)}{\sum_{i=1}^{N} \Phi(\mu_i)} \beta_{\text{log\_price}}$$

where $\phi()$ is the standard normal probability density function.

A simple way to get a value for this elasticity is to set each $\Phi(\mu_i)$ to the sample mean **purchase** rate, which implies $\mu_i = \Phi^{-1}(0.084) \approx -1.38$ for all i. It would be more accurate to plug in the values of the covariates for each member of the population, but setting everyone at the sample mean makes the calculation easier and means that it does not depend on the population size. Plugging $\mu_i = -1.38$ and the posterior mean of $\beta_{\text{log\_price}}$ into the elasticity equation yields

$$\frac{\partial \log E}{\partial \text{log\_price}} \approx \frac{\phi(-1.38)}{\Phi(-1.38)} \beta_{\text{log\_price}} \approx 1.84 \times -5.22 \approx -9.60$$

This is an overestimate, since the estimate for $\beta_{\text{log\_price}}$ is coming from a model that has data only from people who you are pretty sure knew about the price. Many people in the region do not know about a dealership's price changes one way or another, in which case their price elasticity of demand is zero. Suppose only one in ten individuals in the region learn any information about the price, whether through a flier in the mail, word of mouth, or some other means. Then the price elasticity of demand is about –0.96. According to the literature, this seems reasonable. For example, Copeland (2009) finds a price elasticity of demand for GMC pickup trucks of about –1 in a dynamic model and –2 in a simpler static model. So using Rule 10, you can set m = –0.96.

To choose s, it is instructive to start with the original weakly informative prior from the regression model, in that case with $s = 4$. In this case, you have stronger prior information about the likely value of the elasticity, so it is reasonable to tighten down the prior. Setting $s = 0.5$ provides an informative prior, but not too informative. It assumes that there is about a 95% chance that the elasticity is between 0 and –2. A positive elasticity would be very surprising, and although an elasticity of about –2 would be in line with the estimates from the dynamic model dynamic models in Copeland (2009), it is very small relative to the implied elasticity from the probit fit. With these considerations in mind, a reasonable prior for $\beta_{\text{log\_price}}$ is then

$$\beta_{\text{log\_price}} \sim \mathrm{N}(-0.96, 0.5^2)$$

The following code preprocesses the data and fits the model by using the COUNTREG procedure. The resulting classical estimates are shown in Figure 9, and the posterior summaries are shown in Figure 10.

```
data trucksales_count;
   set trucksales;
   log_pop_bachelors = log(pop_bachelors);
   log_pop_below_bachelors = log(pop_below_bachelors);
   log_median_income = log(median_income);
   log_price = log(price);
   log_cost_of_living = log(cost_of_living);
   log_mean_precip = log(mean_precip);
   mean_summer_temp_cs = mean_summer_temp;
   mean_winter_temp_cs = mean_winter_temp;
   keep mean_summer_temp_cs mean_winter_temp_cs area_type
      log_pop_bachelors log_pop_below_bachelors log_median_income log_price
      log_cost_of_living log_mean_precip sales;
run; quit;

proc standard data = trucksales_count mean=0 std=1 out=truckcount_transformed;
   var mean_summer_temp_cs mean_winter_temp_cs;
run; quit;

proc countreg data = truckcount_transformed plots = none;
   class area_type;
   model sales = area_type log_pop_bachelors log_pop_below_bachelors
      log_median_income log_price log_cost_of_living
      log_mean_precip mean_summer_temp_cs mean_winter_temp_cs;
   bayes seed = 56549 ntu = 100 mintune = 20 maxtune = 20 nmc = 10000
      statistics = (summary interval prior);
   prior intercept ~ normal(mean = 8.88, var = 10000);
   prior log_pop_bachelors log_pop_below_bachelors log_median_income
      log_cost_of_living log_mean_precip log_price ~ normal(mean = 0, var = 16);
   prior mean_summer_temp_cs mean_winter_temp_cs
      area_type_rural area_type_sub ~ normal(mean = 0, var = 7.62);
   prior log_price ~ normal(mean = -0.96, var = 0.25);
run; quit;
```

Figure 9 Classical Poisson Regression Estimates of Truck Sales Data

The COUNTREG Procedure

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value	Approx Pr > |t|
Intercept	1	10.560301	6.442817	1.64	0.1012
area_type rural	1	2.001451	0.029271	68.38	<.0001
area_type sub	1	0.999953	0.028986	34.50	<.0001
area_type urban	0	0	.	.	.
log_pop_bachelors	1	0.035539	0.149601	0.24	0.8122
log_pop_below_bachelors	1	-0.064428	0.454199	-0.14	0.8872
log_median_income	1	0.009019	0.028098	0.32	0.7482
log_price	1	-0.598401	0.195036	-3.07	0.0022
log_cost_of_living	1	-0.061695	0.050142	-1.23	0.2185
log_mean_precip	1	0.011931	0.020699	0.58	0.5644
mean_summer_temp_cs	1	-0.006239	0.008432	-0.74	0.4593
mean_winter_temp_cs	1	0.079396	0.008708	9.12	<.0001

Figure 10 Bayesian Poisson Regression Estimates of Truck Sales Data

Posterior Summaries

Parameter	N	Mean	Standard Deviation	Percentiles 25%	50%	75%
Intercept	10000	10.7170	6.2750	6.4443	10.7617	14.8108
area_type_rural	10000	2.0039	0.0308	1.9821	2.0043	2.0242
area_type_sub	10000	1.0032	0.0311	0.9824	1.0019	1.0238
log_pop_bachelors	10000	0.0431	0.1475	-0.0540	0.0435	0.1437
log_pop_below_bachelors	10000	-0.0178	0.4491	-0.3012	-0.0143	0.2869
log_median_income	10000	0.0111	0.0278	-0.00814	0.0103	0.0298
log_price	10000	-0.6738	0.1832	-0.8056	-0.6749	-0.5578
log_cost_of_living	10000	-0.0584	0.0495	-0.0916	-0.0586	-0.0235
log_mean_precip	10000	0.0128	0.0200	-0.00072	0.0130	0.0270
mean_summer_temp_cs	10000	-0.00622	0.00848	-0.0122	-0.00598	-0.00022
mean_winter_temp_cs	10000	0.0792	0.00880	0.0733	0.0796	0.0854

Posterior Intervals

Parameter	Alpha	Equal-Tail Interval		HPD Interval	
Intercept	0.050	-1.5525	23.3088	-1.6171	22.7344
area_type_rural	0.050	1.9439	2.0657	1.9424	2.0634
area_type_sub	0.050	0.9404	1.0634	0.9385	1.0590
log_pop_bachelors	0.050	-0.2460	0.3203	-0.2460	0.3203
log_pop_below_bachelors	0.050	-0.9247	0.8337	-0.9476	0.8073
log_median_income	0.050	-0.0435	0.0659	-0.0414	0.0669
log_price	0.050	-1.0203	-0.3003	-1.0536	-0.3428
log_cost_of_living	0.050	-0.1542	0.0358	-0.1502	0.0365
log_mean_precip	0.050	-0.0269	0.0522	-0.0233	0.0545
mean_summer_temp_cs	0.050	-0.0226	0.0104	-0.0232	0.00946
mean_winter_temp_cs	0.050	0.0612	0.0966	0.0624	0.0972

In this case, the classical estimate of the price elasticity of demand for the truck is around –0.6, whereas the Bayesian estimate with the informative prior is around –0.67. The prior attenuates the estimate toward –1 somewhat, but not a large amount. Because the literature suggests that price elasticities of demand for cars and trucks are typically –1 or –2, and the advertising data suggest that it is about –1, this should improve the quality of your inferences.

The posterior standard deviation of $\beta_{\text{log\_price}}$ is also a bit smaller than the parameter's standard error in the classical estimation. As a result, Bayesian credible intervals are somewhat narrower than classical confidence intervals. From Figure 10, you see that the 95% credible interval for $\beta_{\text{log\_price}}$ is about $(-1.02, -0.30)$, whereas the 95% confidence interval can be computed from Figure 9 as $(-0.99, -0.21)$. If you trust the information in your prior, this should be regarded as a feature of the Bayesian analysis. The credible interval is narrower, to reflect the fact that the prior and data largely agree on the likely values of the parameter, though they do not completely agree.

DISCUSSION

In each of the three examples, several choices need to be made in order to come up with a prior distribution, and typically there is no one right choice. This is an unfortunate reality of statistics: your choices in the data selection, data preprocessing, model selection, and prior selection steps can have a major impact on your results, but in many cases you have only rough guidelines at best. Prior selection is a unique issue in the Bayesian context, but non-Bayesians have their own set of similar issues, such as choosing significance levels for hypothesis tests or minimum detectable effects in power analyses.

There is typically no one best way to make these choices, and indeed there is no one best way to construct priors. No theorem can tell you how to convert vague intuitions that are based on your experience into mathematically precise probability distributions. For other examples of how to navigate these waters, see, for example, Kadane and Wolfson (1998), O'Hagan (1998), or Albert et al. (2012). An alternative approach is to construct "reference" or "objective" priors that satisfy some desirable properties in the case where you have no useful information with which to inform your prior distributions. Two common approaches here are choosing priors that satisfy some intuitive mathematical definition of "ignorance," and choosing priors that result in posterior credible intervals that match the classical confidence intervals. See, for example, Berger and Bernardo (1992) and Berger (2006) for more information about these approaches.

The approach in this paper is to use mathematical tricks, usually transformations, to make quantities in the model easier to think about. Then the analyst can attempt to translate outside knowledge about the problem into a prior distribution on the parameters. By its nature, this process must be ad hoc. This paper presents a number of rules of thumb to guide you through the process, but they should not be taken as general rules that must always apply. The examples illustrate how to apply some of these rules, but the particular choices made there should not be taken as canonical for those examples. There is always room for disagreement on how to best represent prior information, and the goals that you have for fitting the model will further inform those priors. It is good practice to fit the model with several priors that seem to reasonably represent your prior information, whether that information comes from intuitions about the problem or vaguely related data sets. Comparing the results shows you just how sensitive your inferences are to how you translated vague prior information into prior distributions.

Several benefits of taking the Bayesian approach also emerge. First, Bayesian inference enables the analyst to make probability statements about quantities of interest. A 95% credible interval for a parameter is an interval such that the probability that the parameter is in the interval is 0.95, at least according to the model. This is typically much easier for the consumers of model output to interpret than confidence intervals. Second, even weakly informative priors tend to attenuate parameter estimates toward their default values. When set correctly, these priors help protect you from, for example, multiple comparisons issues. Third, when the model and the prior agree on the likely values of a parameter, the resulting posterior distribution is more concentrated around those likely values. If the prior is set well, this is faithfully showing you that your uncertainty about the parameter should decrease. Finally, the process of designing priors that accurately represent your uncertainty, or even a default state of uncertainty, forces you to interrogate your beliefs about the problem and the model. This can lead to new insights into how to think about the problem, potentially resulting in better models, and can often highlight the weak points of the analysis, including the priors. Often this is the greatest benefit of a Bayesian approach: it forces you to think more clearly and be more transparent about the assumptions that go into your analysis.

REFERENCES

Albert, I., Donnet, S., Guihenneuc-Jouyaux, C., Low Choy, S., Mengersen, K. L., and Rousseau, J. (2012). "Combining Expert Opinions in Prior Elicitation." *Bayesian Analysis* 7:503–532.

Berger, J. O. (2006). "The Case for Objective Bayesian Analysis." *Bayesian Analysis* 3:385–402. http://www.stat.cmu.edu/bayesworkshop/2005/berger.pdf.

Berger, J. O., and Bernardo, J. M. (1992). "On the Development of the Reference Prior Method." In *Bayesian Statistics 4: Proceedings of the Fourth Valencia International Meeting*, edited by J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, 35–60. Oxford: Oxford University Press.

Copeland, A. M. (2009). *The Dynamics of Automobile Expenditures*. Federal Reserve Bank of New York Staff Report 394.

Gelman, A. (2006). "Prior Distributions for Variance Parameters in Hierarchical Models." *Bayesian Analysis* 1:515–533.

Gelman, A., and Loken, E. (2014). "The Statistical Crisis in Science." *American Scientist* 102:460–466.

Ghosh, J., Li, Y., and Mitra, R. (2018). "On the Use of Cauchy Prior Distributions for Bayesian Logistic Regression." *Bayesian Analysis* 13:359–383.

Kadane, J. B., and Wolfson, L. J. (1998). "Experiences in Elicitation." *Journal of the Royal Statistical Society, Series D* 47:3–19.

O'Hagan, A. (1979). "On Outlier Rejection Phenomena in Bayes Inference." *Journal of the Royal Statistical Society, Series B* 41:358–367.

O'Hagan, A. (1998). "Eliciting Expert Beliefs in Substantial Practical Applications." *Journal of the Royal Statistical Society, Series D* 47:21–35.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author:

Matthew Simpson
SAS Institute Inc.
SAS Campus Drive
Cary, NC 27513
Matt.Simpson@sas.com