Statistical modeling

April 30, 2018

1 Statistical model

A statistical model will be a mathematical structure used to imitate and approximate, the data generating process. It typically describes relationships among variables while accounting for uncertainty and variability in the data.

1.1 Modelling objectives

1. **Quantify uncertainty** - Tackled with theory of probability
2. **Inference** - What is the behavior of the whole population
3. **Measure support for hypothesis** - Is this strong enough evidence to support or validate the experts claim?
4. **Prediction**

1.2 Statisticam modelling process

1. Understand the problem
2. Plan and collect data
3. Explore the data
4. Postulate model (choose the model)
5. Fit model
6. Check the model
7. Iterate (steps: 4-6)
8. Use the model

1.3 Questions

Question 1

Which objective of statistical modeling is best illustrated by the following example?

You fit a linear regression of monthly stock values for your company. You use the estimates and recent stock history to calculate a forecast of the stock's value for the next three months.

a. Quantify uncertainty

b. Inference

c. Hypothesis testing

d. Prediction

Answer 1

d.

Question 2

Which objective of statistical modeling is best illustrated by the following example?

A biologist proposes a treatment to decrease genetic variation in plant size. She conducts an experiment and asks you (the statistician) to analyze the data to conclude whether a 10% decrease in variation has occurred.

a. Quantify uncertainty

b. Inference

c. Hypothesis testing

d. Prediction

Answer 2

c.

Question 3

Which objective of statistical modeling is best illustrated by the following example?

The same biologist form the previous question asks you how many experiments would be necessary to have a 95% chance at detecting a 10% decrease in plant variation.

a. Quantify uncertainty

b. Inference

c. Hypothesis testing

d. Prediction

Answer 3

a.

Question 4
Which of the following scenarios best illustrates the statistical modeling objective of inference?

a. A model inputs academic performance of 1000 students and predicts which student will be valedictorian after another year of school.

b. A social scientist collects data and detects positive correlation between sleep deprivation and traffic accidents.

c. A natural language processing algorithm analyzes the first four words of a sentence and provides words to complete the sentence.

d. A venture capitalist uses data about several companies to build a model and makes recommendations about which company to invest in next based on growth forecasts.

Answer 4
b
Question 5
Which step in the statistical modeling cycle was not followed in the following scenario?

Susan gathers data recording heights of children and fits a linear regression predicting height from age. To her surprise, the model does not predict well the heights for ages 14-17 (because the growth rate changes with age), both for children included in the original data as well as other children outside the model training data.

a. Use the model

b. Explore the data

c. Plan and properly collect relevant data

d. Fit the model

Answer 5
b
Question 6
Which of the following is a possible consequence of failure to plan and properly collect relevant data?

a. You will not produce enough data to make conclusions with a sufficient degree of confidence.

b. Your selected model will not be able to fit the data.

c. You may not be able to visually explore the data.

d. Your analysis may produce incomplete or misleading results.

Answer 6

d.

Xie operates a bakery and wants to use a statistical model to determine how many loaves of bread he should bake each day in preparation for weekday lunch hours. He decides to fit a Poisson model to count the demand for bread. He selects two weeks which have typical business, and for those two weeks, counts how many loaves are sold during the lunch hour each day. He fits the model, which estimates that the daily demand averages 22.3 loaves.

Over the next month, Xie bakes 23 loaves each day, but is disappointed to find that on most days he has excess bread and on a few days (usually Mondays), he runs out of loaves early.

Question 7

Which of the following steps of the modeling process did Xie skip?

a. Understand the problem

b. Postulate a model

c. Fit the model

d. Check the model and iterate

e. Use the model

Answer 7

c

Question 8

What might you recommend Xie do next to fix this omission and improve his predictive performance?

a. Abandon his statistical modeling initiative.

b. Collect three more weeks of data from his bakery and other bakeries throughout the city. Re-fit the same model to the extra data and follow the results based on more data.

c. Plot daily demand and model predictions against the day of the week to check for patterns that may account for the extra variability. Fit and check a new model which accounts for this.

d. Trust the current model and continue to produce 23 loaves daily, since in the long-run average, his error is zero.

Answer 8

c

2 Bayesian Modeling

2.1 Components of Bayesian Models

Let:

heights $n = 15$ men

$$y_i = \mu + \epsilon_i, \epsilon_i \overset{iid}{\sim} N(0, \sigma^2), i = 1, ..., n$$

$$y_i \overset{iid}{\sim} N(\mu, \sigma^2)$$

So far, this model is the same for frequentists and Bayesians.

The frequentist approach to fitting this model right here. Would be to consider μ and σ to be fixed but unknown constants, and then we would estimate them. To calculate our uncertainty in those estimates. A frequentist approach would consider how much the estimates of μ and σ might change. If we were to repeat the sampling process and obtain another sample of 15 men, over, and over.

The Bayesian approach, the one we're going to take here. Tackles our uncertainty in μ and σ^2 with probability directly. By treating them as random variables with their own probability distributions.

The three primary components of Bayesian models that we often work with are:

1) **likelihood** ($p(y|\theta)$)

2) **prior** ($p(\theta)$the probability distribution that characterizes our uncertainty with the parameter θ.)

3) **posterior** ($p(\theta|y) = \dfrac{p(\theta,y)}{p(y)} = \dfrac{p(\theta,y)}{\int p(\theta,y)d\theta} = \dfrac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta}$).

2.2 Model Specification

Before fitting any model we first need to specify all of its components. One convenient way to do this is to write down the hierarchical form of the model. By hierarchy, we mean that the model is specified in steps or in layers.

1^{rst} **level - likelihood**

$$y_i|\mu,\sigma^2 \overset{iid}{\sim} N(\mu,\sigma^2)i = i,...,n$$

** 2^{nd} level - the prior distribution from μ and σ^2 **

$$P(\mu,\sigma^2) = p(\mu)p(\sigma^2)$$

For now we're going to say that they're independent priors. We can assume independents in the prior and still get dependents in the posterior distribution.

The conjugate prior for mu, if we know the value of σ^2, is a normal distribution, and the conjugate prior for σ^2 when μ is known is the inverse gamma distribution.

** 3^{rd} level - Prior distribution for μ and σ^2**

$$\mu \sim N(\mu_0,\sigma_0^2)$$
$$\sigma^2 \sim IG(\nu,\beta_0)$$

A useful way to write the model is as a graphical representation which illustrates the dependences among the parameters.

We start from the priors (μ and σ^2) and finish with the likelihood. We denote the random variables that follow their own distribution by a single cycle and the observed values as a double cycle. The square denotes that the y's come from the same distribution across all values of i.

2.3 Posterior derivation

So far, we've only drawn the model with two levels. But in reality, there's nothing that'll stop us from adding more layers. Let's go through the previous example and illustrate it as an hierchical model. One reason we might do this is if the data are hierarchically organized so that the observations are naturally grouped together.

$$y_i|\mu,\sigma^2 \overset{iid}{\sim} N(\mu,\sigma^2) i = i, ..., n$$

$$\mu|\sigma^2 \sim N(\mu_0, \frac{\sigma^2}{w_0})$$

$$\sigma^2 \sim IG(\nu, \beta_0)$$

This can be illustrated as :

Once we have a model specification, we can write out what the full posterior distribution for all the parameters given the data looks like.

Remember that the numerator in Bayes' theorem is the joint distribution of all random quantities, all the nodes in this graphical representation over here from all of the layers. So for this model that we have right here, we have a joint distribution that'll look like this:

$$\begin{aligned} P(y_1, ..., y_n, \mu, \sigma^2) &= P(y_1, ..., y_n|\mu, \sigma^2)p(\mu|\sigma^2)p(\sigma^2) \\ &= \prod_{i=1}^{n}[N(y_i|\mu,\sigma^2)] \times N(\mu|\mu_0, \frac{\sigma^2}{w_0}) \times IG(\sigma^2|\nu_0, \beta_0) \\ &\propto p(\mu, \sigma^2|y_1, ..., y_n) \end{aligned}$$

Please note that the distribution abbreviation stands for the density function

But how did we derive that the joint distribution of everything is proportional to the

posterior distribution of μ and σ^2, given all of the data?

Remember that:

$$\begin{aligned} p(\theta|y) &= \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta} \\ &\propto p(y|\theta)p(\theta) \end{aligned}$$

$$\text{since } \int p(y|\theta)p(\theta)d\theta \text{ is constant}$$

The only thing missing in this expression right here is just some constant number that causes the expression to integrate to 1. If we can recognize this expression as being proportional to a common distribution, then our work is done, and we know what our posterior distribution is.

If we do not use conjugate priors or if the models are more complicated, then the posterior distribution will not have a standard form that we can recognize.

2.4 Non-conjugate models

Let's go over a couple examples of models that don't have clean posterior distributions.

** Example 1**

Suppose we have values that represent the percentage change in total personnel from last year to this year for, we'll say, ten companies.

These companies come from a particular industry. We're going to assume for now:

$$y_i|\mu \overset{iid}{\sim} N(\mu, 1)\text{nknown mean could represent growth for this particular industry}$$

$$\text{Let: } \mu \sim t(0,1,1)$$

Recall that the posterior distribution of mu is proportional to the likelihood times the prior. Let's write the expression for that in this model.

$$P(\mu|y_1, ..., y_n) \propto \prod_{i=1}^{n}[\frac{1}{\sqrt{2\pi}} exp(-\tfrac{1}{2}(y_i - \mu)^2)] \frac{1}{\pi(1+\mu^2)}$$

$$\propto exp[-\frac{1}{2}\sum_{i=1}^{n}(y_i - \mu)^2] \frac{1}{1+\mu^2}$$

$$\propto exp[-\frac{1}{2}(\sum_{i=1}^{n} y_i^2 - 2\mu \sum_{i=1}^{n} y_i + n\mu^2] \frac{1}{1+\mu^2}$$

$$\propto \frac{exp[n(\bar{y}\mu - \frac{\mu^2}{2})]}{1+\mu^2}$$

Example 2

$$y_i|\mu, \sigma^2 \overset{iid}{\sim} N(\mu, \sigma^2) i = i, ..., n$$

$$\mu|\sigma^2 \sim N(\mu_0, \sigma_0^2)$$

$$\sigma^2 \sim IG(\nu, \beta_0)$$

We saw earlier that if you include sigma squared in the prior for mu, and use the hierarchical model that we presented earlier, that model would be conjugate and have a closed form solution.

However, in the more general case that we have right here, the posterior distribution does not appear as a distribution that we can simulate or integrate. Challenging posterior distributions like these ones and most others that we'll encounter later on kept Bayesian in methods from entering the main stream of statistics for many years. Since only the simplest problems were tractable.

However, computational methods invented in the 1950's, and implemented by statisticians decades later, revolutionized the field. We do have the ability to simulate from the posterior distributions in this lesson as well as for many other more complicated models.

What is the major challenge do we face with both of the models introduced in this segment?

We have the posterior distribution up to a normalizing constant, but we are unable to integrate it to obtain important quantities, such as the posterior mean or probability intervals.

In low dimensional problems with only a few parameters we can resort to numerical methods for integration, but this solution only works for a narrow set of models.