STAT 330: Problem Set 2

Freddie Hall, Mariah Boudreau, and Erik Weis

September 30, 2021

1.1 Where's the line, likelihood?

In the model definition below, which line describes the likelihood?

$$\begin{gathered} X_i \overset{i.i.d.}{\sim} Normal(\mu, \sigma^2) \\ \mu \sim Normal(0, 1) \\ \sigma^2 \sim Uniform(-5, 5) \end{gathered}$$

Solution. The first line is the likelihood as it shows the relationship between the data and the model parameters μ and σ^2.

1.2 Consider the following partial model for the data $(W_i, X_i, Y_i) i = 1, ..., N$.

$$\begin{gathered} Y_i \overset{i.i.d.}{\sim} Normal(\mu_i, \sigma^2) \\ \mu_i \sim \alpha + \beta_1 W_i + \beta_2 X_i \end{gathered}$$

How many priors would we need to add to complete this model?

Solution. We would need to add four more priors to this model, a prior for σ^2, α, β_1, and β_2.

1.3 Consider the model

$$\begin{gathered} X_i \sim Bernoulli(p) \\ p \sim Beta(1, 1) \end{gathered}$$

Where each data point $X_i \in \{H, T\}$ is the outcome of a coin toss, and the parameter is the bias p of the coin. What is the support of the prior-predictive distribution? What is the support of the posterior-predictive distribution?

Solution. The support of the prior-predictive distribution is same as the possible values of the data X_i, which is dictated by the model. In this case, those values are H and T. The support of the posterior-predictive is also H and T, and again represent all the possible values X_i can take, as dictated by the model.

1.4 Consider a model with a single parameter θ on which we place a uniform prior $P(\theta) \sim 1$. Why is the MAP estimator of θ equivalent to the maximum likelihood estimator (MLE) typically used in "classical statistics?"

Solution. For any posterior distribution,

$$P(\theta|X) \propto P(X|\theta)P(\theta).$$

For a uniform prior means the prior distribution is constant. Hence,

$$P(\theta|X) \propto P(X|\theta).$$

Because these two functions are proportional, the max of the likelihood (MLE) and the max of the posterior (MAP) will occur for the same value θ. More formally, the derivative of the likelihood and posterior will be zero for the same value(s) of θ.

2 Theoretical walkthrough

Assume that you have the posterior distribution $P(\theta|X)$ for a single parameter θ with support on the reals $\mathbb{R} = (-\infty, \infty)$.

We will show the estimator $\hat{u}$ that minimizes the expectations of the loss function

$$f(\theta; u) = \begin{cases} k_1(u - \theta) & \text{if } \theta < u, \\ k_2(\theta - u) & \text{if } \theta > u \end{cases}$$

is the $k_2/(k_1 + k_2)$ fractile of the posterior distribution, i.e., the value $\hat{u}$ of θ below which a fraction $k_2/(k_1 + k_2)$ of the posterior probability mass resides.

2.1 Write the expectation $\mathbb{E}[f(\theta; u))]$ of $f(\theta; u)$ over the posterior distribution $P(\theta|X)$.

Solution.

$$\begin{aligned} \mathbb{E}[f(\theta; u))] &= \int_{-\infty}^{\infty} f(\theta; u)P(\theta|X)d\theta \\ &= \int_{u}^{\infty} k_2(\theta - u)P(\theta|X)d\theta + \int_{-\infty}^{u} k_1(u - \theta)P(\theta|X)d\theta \end{aligned}$$

2.2 Using the Leibniz integral rule for a function $h(u|\theta)$

$$\frac{d}{du}\int_{a(u)}^{b(u)} h(u, \theta)d\theta = h(u, b(u))\frac{d}{du}b(u) - h(u, a(u))\frac{d}{du}a(u) + \int_{a(u)}^{b(u)} \frac{\partial}{\partial u}h(u, \theta)d\theta,$$

compute the derivative of the expected loss with respect to the parameter u. Setting this derivative to 0 will give you an implicit equation for the estimator $\hat{u}$ that minimizes this loss,

$$\frac{d}{du}\mathbb{E}[f(\theta, u)]|_{u=\hat{u}} = 0.$$

Solution. We can start by rewriting the expected value of the loss function in a more useful form.

$$\begin{aligned}\mathbb{E}[f(\theta,u) &= \int_u^\infty k_2(\theta-u)P(\theta|X)d\theta + \int_{-\infty}^u k_1(u-\theta)P(\theta|X)d\theta \\ &= \int_u^\infty h_1(\theta,u)d\theta + \int_{-\infty}^u h_2(\theta,u)d\theta \\ &= \lim_{a\to\infty}\left[\int_u^a h_1(\theta,u)d\theta + \int_{-a}^u h_2(\theta,u)d\theta\right]\end{aligned}$$

Now, taking the derivative with respect to u, we get

$$\begin{aligned}\frac{d}{du}\mathbb{E}[f(\theta,u)] &= \frac{d}{du}\lim_{a\to\infty}\left[\int_u^a h_1(\theta,u)d\theta + \int_{-a}^u h_2(\theta,u)d\theta\right] \\ &= \lim_{a\to\infty}\Bigg[h_1(u,u)\frac{d}{du}(u) + h_1(-a,u)\frac{d}{du}(a) + \int_{-a}^u k_1 P(\theta|X)d\theta \\ &\qquad + h_2(u,a)\frac{d}{du}(a) + h_2(u,u)\frac{d}{du}(u) + \int_u^a -k_2P(\theta|X)d\theta\Bigg]\end{aligned}$$

We note that because the loss function is zero if $\theta = u$, $h_1(u,u) = h_2(u,u) = 0$. Additionally, $\frac{d}{du}(a) = 0$ because a is a constant. Therefore, all the non-integral terms are zero, leaving the expression

$$\lim_{a\to\infty}\left[\int_{-a}^u k_1P(\theta|X)d\theta - \int_u^a k_2P(\theta|X)d\theta\right]$$

Taking the limit, evaluating the expression at $u = \hat{u}$, and setting equal to zero, we obtain an equation for the minimum of the expected lost $\hat{u}$.

$$\int_{-\infty}^{\hat{u}} k_1P(\theta|X)d\theta - \int_{\hat{u}}^\infty k_2P(\theta|X)d\theta = 0$$

2.3 Re-express the equation found at the previous step in terms of the distribution function of $P(\theta|X)$

$$F_\theta(t) = \int_{-\infty}^t P(\theta|X)d\theta$$

Then show

$$F_\theta(\hat{u}) = \frac{k_2}{k_1+k_2}$$

Solution. Let

$$F_\theta(\hat{u}) = \int_{-\infty}^{\hat{u}} P(\theta|X)d\theta$$
$$F_\theta(\hat{u}) = 1 - \int_{\hat{u}}^{\infty} P(\theta|X)d\theta$$

Making these substitutions gives

$$\begin{aligned} 0 &= \int_{-\infty}^{\hat{u}} k_1 P(\theta|X)d\theta - \int_{\hat{u}}^{\infty} k_2 P(\theta|X)d\theta \\ &= k_1 F_\theta(u) - k_2(1 - F_\theta(\hat{u})) \\ &= k_1 F_\theta(\hat{u}) - k_2 + k_2 F_\theta(\hat{u}) \end{aligned}$$

Solving this equation

$$k_2 = (k_1 + k_2)F_\theta(\hat{u})$$
$$F_\theta(\hat{u}) = \frac{k_2}{k_1 + k_2}$$

2.4 Explain why the equation above tells us that the optimal estimator $\hat{u}$ is the $k_2/(k_1+k_2)$ using your own words, a sketch, or both.

Solution. Given the substitution above of,

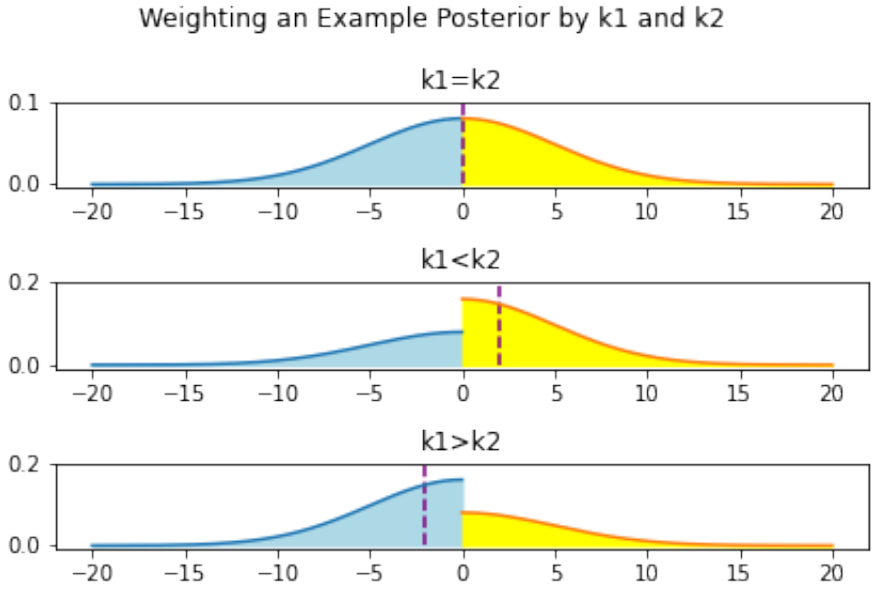
$$0 = \int_{-\infty}^{\hat{u}} k_1 P(\theta|X)d\theta - \int_{\hat{u}}^{\infty} k_2 P(\theta|X)d\theta$$

We can achieve this expression then

$$k_1 \int_{-\infty}^{\hat{u}} P(\theta|X)d\theta = k_2 \int_{\hat{u}}^{\infty} P(\theta|X)d\theta$$

This equivalence shows that the posterior is being weighted by the variables k_1 and k_2. If given $k_1 = k_2$, then each half of the posterior distribution is equivalent. Therefore, we can define the proportion of the area under the curve for which $\hat{u}$ is the optimal estimator as the $\frac{k_2}{(k_1+k_2)}$, since k_2 is the constant associated with $\hat{u}$ being less than θ. The goal is given various values for the constants, the value of $\hat{u}$ needs to adjust to conserve the equivalence given above.

Shown visually with a normal posterior as an example,



The equation below is also representative of probability mass below $\hat{u}$. We see that $\frac{k_2}{k_1+k_2}$ is just a fraction of the total probability mass.

$$\int_{-\infty}^{\hat{u}} P(\theta|X)d\theta = \frac{k_2}{(k_2+k_1)} \int_{-\infty}^{\infty} P(\theta|X)d\theta$$

We've also plotted the function $g(\theta, u) = f(\theta, u)P(\theta|X)$ interactively (see associated jupyter notebook).

2.5 Using the Leibniz integral rule again, show that the *second* derivative of the expected loss evaluated at $\hat{u}$ is positive, i.e., that

$$\frac{d^2}{du^2}\mathbb{E}[f(\theta, u)]\bigg|_{u=\hat{u}} > 0.$$

Explain why this inequality tells us that $\hat{u}$ is a minimum of the expected loss, instead of a maximum or an inflection point(where the first derivative is also zero).

Solution. Remember the first derivative the expectation of the loss function is

$$\lim_{a\to\infty}\Big[\int_{-a}^{\hat{u}} k_1 P(\theta|X)d\theta - \int_{\hat{u}}^{a} k_2 P(\theta|X)d\theta\Big]$$

Taking the second derivative will require using the Leibniz integral rule, giving,

$$\begin{aligned}
\frac{d^2}{du^2}\mathbb{E}[f(\theta,u)]\bigg|_{u=\hat{u}} &= \frac{d}{d\hat{u}}\lim_{a\to\infty}\Big[\int_{-a}^{\hat{u}} k_1 P(\theta|X)d\theta - \int_{\hat{u}}^{a} k_2 P(\theta|X)d\theta\Big] \\
&= \lim_{a\to\infty}\Big[k_1 P(\hat{u}|X)\frac{d}{d\hat{u}}\hat{u} - k_1 P(-a|X)\frac{d}{d\hat{u}}(-a) + \int_{-a}^{\hat{u}}\frac{d}{d\hat{u}}k_1 P(\theta|X)d\theta \\
&\qquad - k_2 P(a|X)\frac{d}{d\hat{u}}(a) + k_2 P(\hat{u}|X)\frac{d}{d\hat{u}}\hat{u} + \int_{\hat{u}}^{a}\frac{d}{d\hat{u}}k_2 P(\theta|X)d\theta\Big] \\
&= k_1 P(\hat{u}|X) + k_2 P(\hat{u}|X) > 0
\end{aligned}$$

The second derivative gives the above expression, which we know is greater than 0 because both k_1 and k_2 are both non-negative integers being multiplied by a posterior distribution, which is by definition non-negative. This inequality tells us that $\hat{u}$ is a minimum of the expected loss because the second derivative deals with concavity, and to have a positive concavity means that the shape is concave up, which is correlates with a minimum value rather than a maximum value.

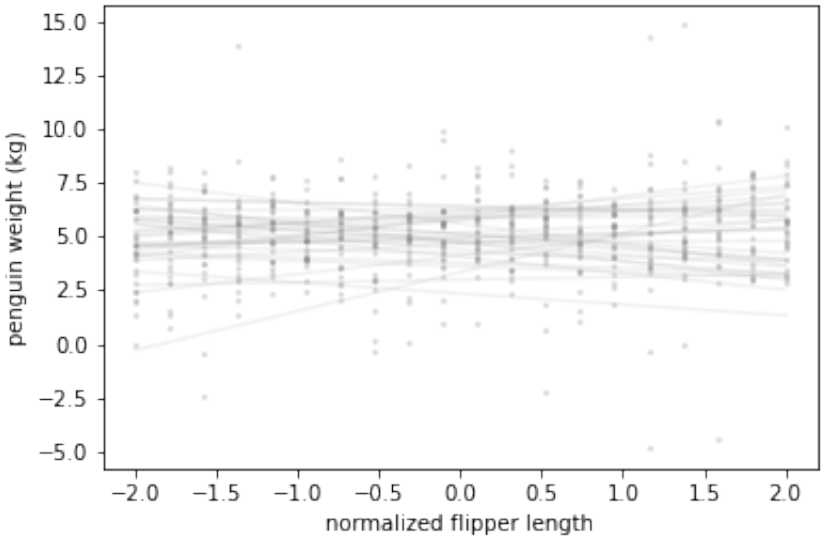
3.1 Load the Penguin data set and normalize the flipper length so that the mean of this column equals 0 and the standard deviation of the sample is 1. Do not normalize the mass, but convert it to kilograms.

3.2 Let Y_i be the body mass of penguin i and X_i be its normalized flipper length. For the model

$$\begin{aligned}
Y_i &\sim \text{Normal}(\mu_i, \sigma^2) \\
\mu_i &= \alpha. + \beta X_i
\end{aligned}$$

formulate a prior for α, β, and σ^2

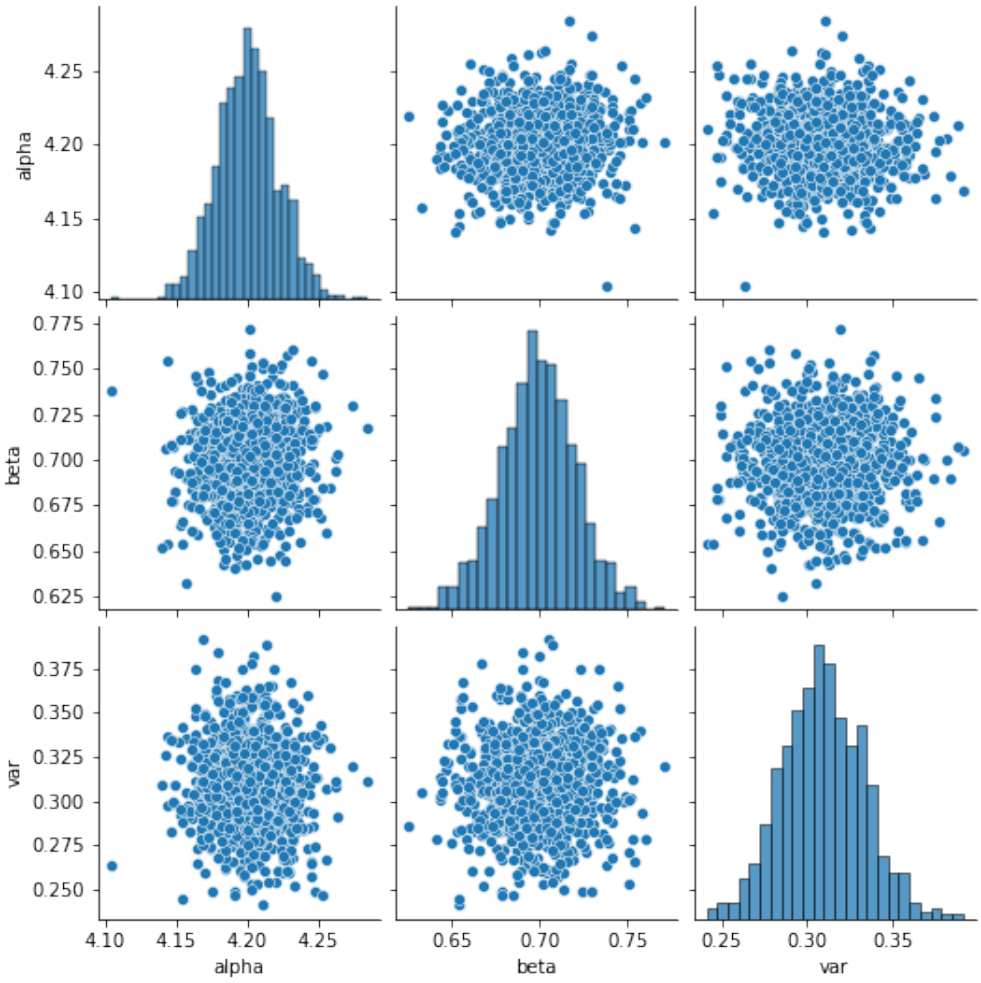
3.3 Argue that your prior is reasonable using samples from the prior-predictive distribution. Look both at predicted regressions lines $\mu = \alpha + \beta x$ and predicted data sets Y.



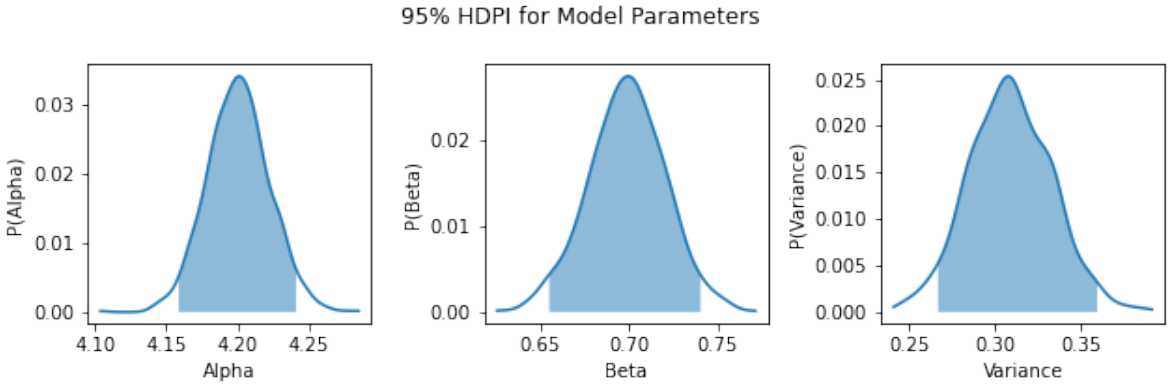
3.4 Write an expression for the posterior distribution of the model. Compute this posterior distribution with the grid or quadratic approximation and report the MAP estimators.

3.5 Generate samples from the posterior distribution and use these samples to show the marginal distribution of each parameter. Compute percentile or highest density posterior intervals for each parameter.

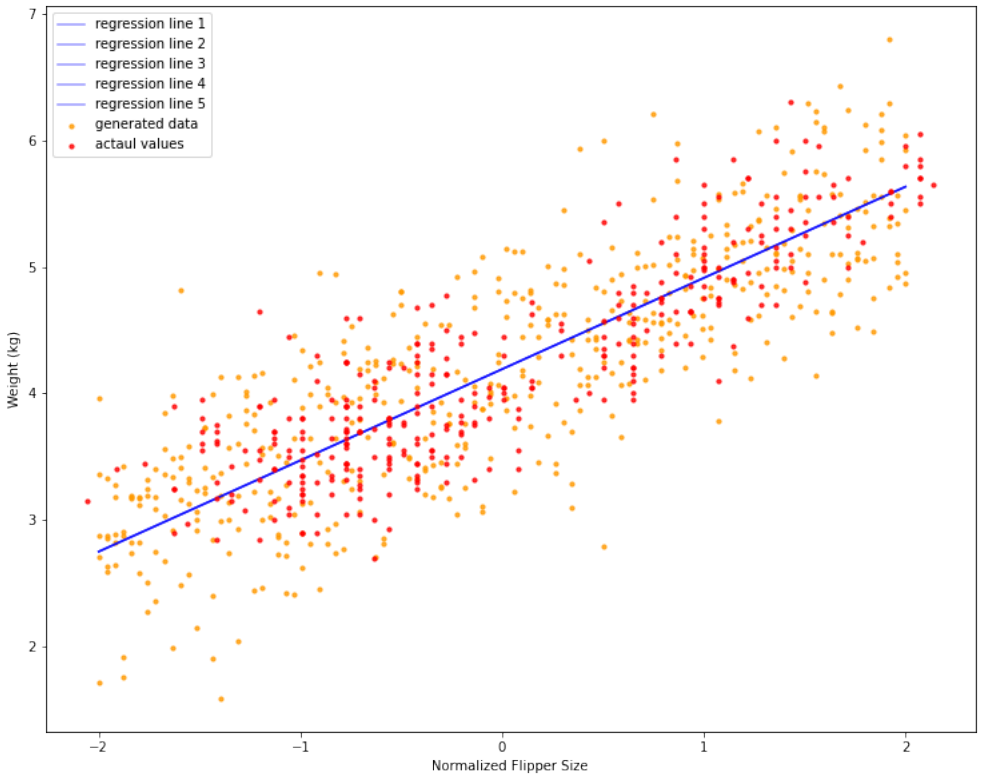
The marginal distributions of the parameters α, β, and σ^2 are shown in the figure below.



We also show the highest density posterior intervals (HDPI) using a kernel density estimator function.



3.6 Provide a visual verification that the estimated model fits the data well, using posterior predictive checks. Show both the estimated regression lines and data simulated from the posterior predictive distribution.


regression line 1
regression line 2
regression line 3
regression line 4
regression line 5
generated data
actaul values
Weight (kg)
Normalized Flipper Size
7
6
5
4
3
2
−2
−1
0
1
2