STAT 535 Homework 4 Example Solutions – Spring 2024

1) Exercise 9.1:
a) The normal prior is reasonable since β_0 and β_1 are continuous quantities that can take any real value, which matches the support of the normal.
b) σ can only take positive values, so the support of the normal does not match this.
c) From the book: weakly informative priors reflect general prior uncertainty about the model parameters, across a set of sensible parameter values. However, a vague prior might be so vague that it puts weight on *non-sensible* parameter values.

2) Exercise 9.3:
a) β_0 is the expected height (in cm) of a zero-month-old (newborn) kangaroo, and β_1 is the change in expected height (in cm) for each one-month increase in the kangaroo's age. I would expect *a priori* that β_1 is positive.
b) β_0 is the expected number of followers for a data scientist with zero commits in the past week, and β_1 is the change in number of followers for each one-commit increase in the number of commits. I would expect *a priori* that β_1 is positive.
c) β_0 is the expected number of visitors on a zero-rainfall (sunny) day, and β_1 is the change in expected number of visitors for each one-inch increase in rainfall. I would expect *a priori* that β_1 is negative.
d) β_0 is the expected number of hours of Netflix for a person who sleeps zero hours(!), and β_1 is the change in expected number of hours of Netflix for each one-hour increase in sleeping. I would expect *a priori* that β_1 is negative.

For problems 3-6: See R code on course web page for details of the solutions.

3) a) My **D** is a 2x2 diagonal matrix with (2,2) along the diagonals, implying that trace($\mathbf{D}^{-1}$) = 1. My prior on $\mathbf{X}^{\sim}\boldsymbol{\beta}|\tau$ is MVN($\mathbf{y}^{\sim}$, $\tau^{-1}\mathbf{D}$). Based on the hypothetical observations, the prior mean on $\boldsymbol{\beta}$ is (45, 1.75)'.
b) The prior on τ is gamma with parameters a = 0.1 and b = 236.5.
c) A posterior point estimate for τ is about 0.0075. A posterior point estimate for σ^2 is about 127.8 or 134.0, depending on whether the posterior mean or median is used.
d) Summary of point estimates and 95% credible intervals for the β's:

```
          0.025 Quantile 0.5 Quantile 0.975 Quantile
intercept       27.71439     47.23493      66.618756
x1               1.46947      1.88009       2.292478
```

The estimated regression equation is
price^ = 47.23 + 1.88*income
For a person with income $60,000 (income=60), we predict a sales price of 47.23 + 1.88*60 = 160.03 thousand dollars, or $160,030.

4) a)

```
     mean Std. Error 0.025 Quantile 0.5 Quantile 0.975 Quantile
   0.0155     0.5026        -1.0557       0.0155         1.0868
x1 0.0027     0.0023        -0.0022       0.0027         0.0075
x2 0.0535     0.0129         0.0260       0.0535         0.0811
x3 0.0147     0.0145        -0.0162       0.0147         0.0455
```

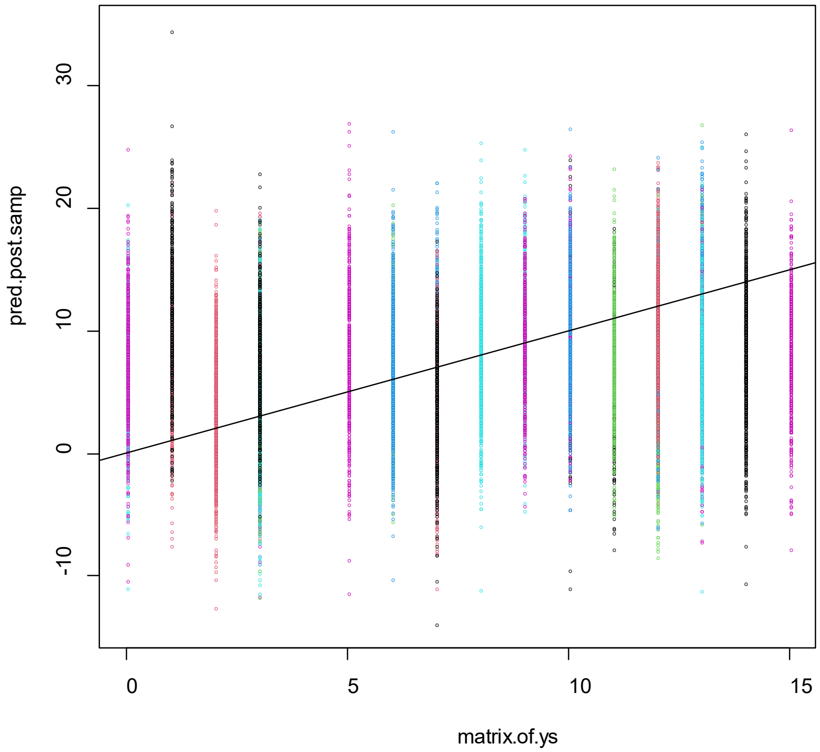
The estimated regression equation is
price^ = 0.0155 + 0.0027*calories + 0.0535*protein + 0.0147*fat

b) Note that the credible intervals for β_1 and β_3 contain 0, but the credible interval for β_2 does not contain 0. So the variable X2=protein has a high posterior probability of having a strong marginal effect on price.
c) The model that has the highest posterior probability is the model with an intercept and "protein" as the only predictor.
d) Among sensible models, the model that has the highest posterior probability by far is still the model with an intercept and "protein" as the only predictor.

5) a) Answers will vary, but from a noninformative prior analysis, an estimated regression equation is:
Sugar^ = 11.975 + 0.0085*Sodium – 2.114*Fiber – 0.464*Carbohydrates + 0.0512*Potassium

Answers will vary slightly if you used a subjective Bayesian prior. The example R code on the course webpage provides a possible subjective prior analysis using the `stan_glm` function.

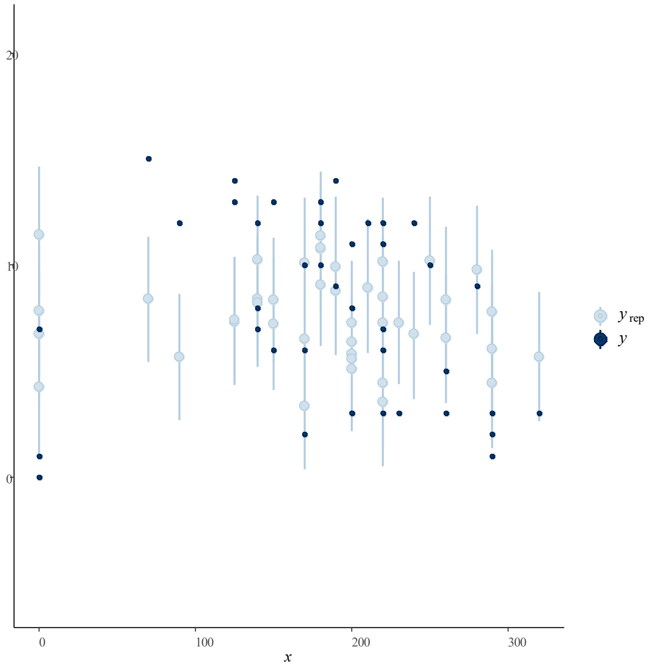
b) Again from the noninformative analysis:



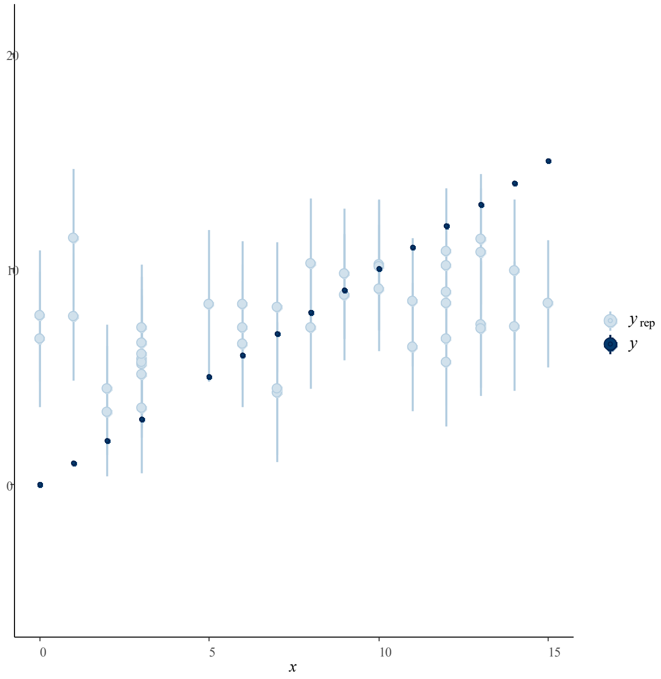
Most of the y-values tend to fall in the middle of the posterior predictive distribution for the respective observations. The cereals with low sugar values of 0 (Puffed Wheat, Puffed Rice) and 1 (Quaker Oatmeal, Cheerios) tend to have y-values that are a little small relative to the predictive distribution. And the high-sugar cereal (Smacks) that a Sugar value that is in the upper tail of the predictive distribution. So there is a little lack of fit for the more extreme observation, but overall the model fit is good.

The rstanarm function produces some other model fit tools. Here is one plot to judge predictive accuracy:

With “Sodium” on the x-axis:



With “Sugar” itself on the x-axis:



c) For the noninformative analysis: The predicted sugar amount for a cereal having sodium=140, fiber=3.5, carbohydrates=14, and potassium=90 is 3.45 grams. The 90% prediction interval is (-4.3, 11.3), but Sugar values cannot be negative, so we can report it as [0, 11.3).

6) Answers will vary. Using an approach that fits models using `stan_glm` and considers only first-order predictors as candidates, the best models based on the ELPD criterion are the ones with
Fiber, carbohydrates, and potassium (ELPD = -127.08)
Fiber and potassium (ELPD = -127.14)
Carbohydrates (ELPD = -127.33)