Data 102 (Fall 2022)
Discussion Worksheet 6
09/30/2022

Names:

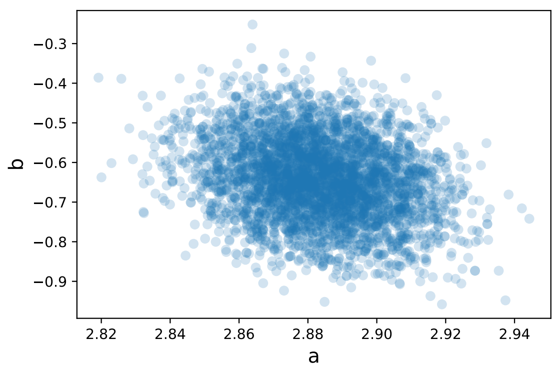
1. You are fitting a Bayesian linear regression model to data points $(x_1, y_1), \ldots, (x_n, y_n)$ where $n = 100$. You posit a distribution

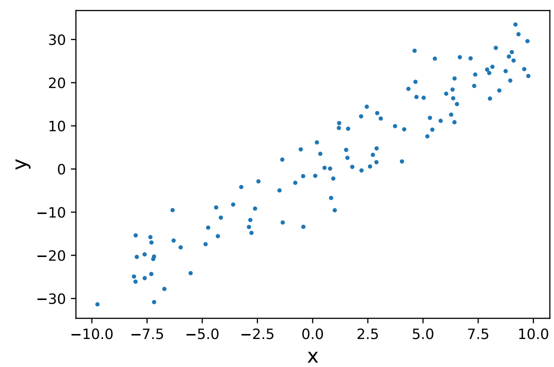
$$Y \mid X \sim \text{Normal}(aX + b, 1),$$

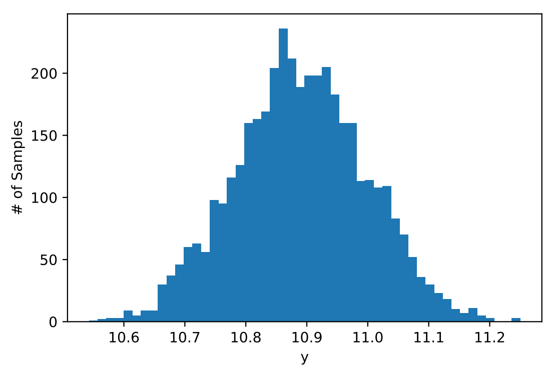
that is, Y is approximately $aX + b$, with an error term that is normally-distributed with standard deviation 1.

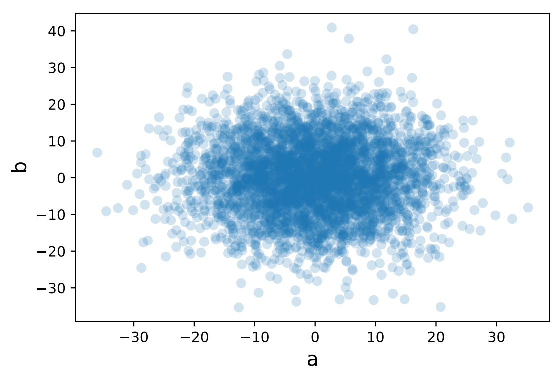
For the Bayesian prior, you place independent normal distributions with standard deviation 10 for both a and b.

(a) The plots below show, in some order, (1) the data, (2) samples from the prior over (a, b), (3) samples from the posterior over (a, b), and (4) the posterior predictive distribution at $x = 4$. Match the plots to these quantities.









Solution: L. to R. : posterior, data, posterior predictive distribution, prior

(b) Does the model pass the posterior predictive check? Why or why not?

Solution: Comparing the ppd samples to the data, the ppd samples overshoot the true value of y corresponding to $x = 4$ and they suggest much smaller variance than the original data.

2. You have a coin with unknown probability π of being heads. You observe 100 independent flips from the coin, of which 70 are heads and 30 are tails.

(a) Write down a prior such that 0.7 is not in the 95% credible interval for π.

Solution:

$$\pi \sim \text{Beta}(1, 41) \tag{1}$$

The posterior concentrates around 0.5, so 0.7 will fall in the tail, outside a 95% credible interval.

(b) Write down a prior such that 0.7 is in the 95% credible interval for π.

Solution:

$$\pi \sim \text{Beta}(1, 1) \tag{2}$$

The posterior concentrates around 0.7 (given form of this prior, posterior is heavily influenced by the data).

(c) Write down a prior such that 0.5 is in the 95% credible interval for π.

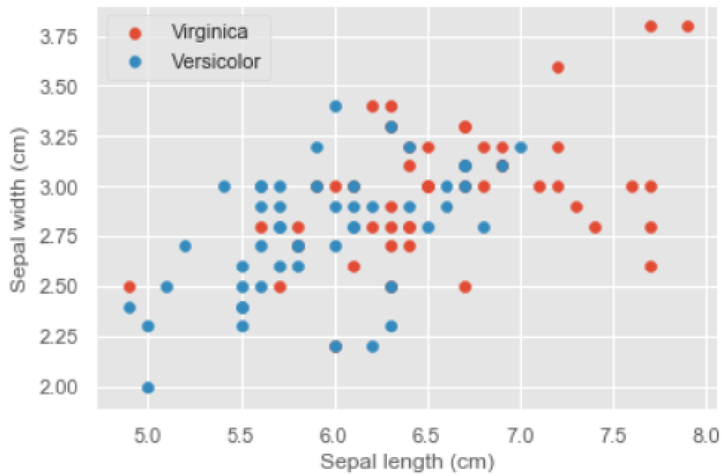
Solution:

$$\pi \sim \text{Beta}(30, 70) \tag{3}$$

The posterior concentrates around 0.5.

3. We have a sample of 100 iris flowers, and measure their sepal length, sepal width, and petal width (sepals are the small, green growths at the base of a flower). The response labels are whether they belong to the *Virginica* species (1) or *Versicolor* species (0).

   The data are represented in the following plot:



   (a) Let's say we first fit a Logistic regression model to predict the iris species, using only the sepal features. After fitting this model, you observe the following output:

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                species   No. Observations:                  100
Model:                            GLM   Df Residuals:                       97
Model Family:                Binomial   Df Model:                            2
Link Function:                  logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -55.163
Date:                Mon, 22 Feb 2021   Deviance:                       110.33
Time:                        00:47:22   Pearson chi2:                     100.
No. Iterations:                     4
Covariance Type:            nonrobust
================================================================================
                  coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
const         -13.0460      3.097     -4.212      0.000     -19.117      -6.975
sepal_width     0.4047      0.863      0.469      0.639      -1.286       2.096
sepal_length    1.9024      0.517      3.680      0.000       0.889       2.916
================================================================================
```

   Under this model, if we increase sepal length by 1cm, how much does this increase the odds of *Virginica* relative to *Versicolor*?

   Solution: Holding sepal width constant, every 1 cm increase in sepal length increases the log odds ratio of the iris belonging to Virginica (rather than Versicolor) by 1.9. Therefore, the odds increase by a factor of $\exp(1.9) = 6.69$.

   (b) We now build another logistic model which additionally includes petal width as a feature. You are presented with the following summary output:

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                species   No. Observations:                  100
Model:                            GLM   Df Residuals:                       96
Model Family:                Binomial   Df Model:                            3
Link Function:                  logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -12.951
Date:                Mon, 22 Feb 2021   Deviance:                       25.902
Time:                        00:47:22   Pearson chi2:                     32.6
No. Iterations:                     8
Covariance Type:            nonrobust
================================================================================
                   coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
const          -20.2873      8.055     -2.519      0.012     -36.075      -4.499
sepal_width     -4.8233      2.097     -2.300      0.021      -8.933      -0.714
sepal_length     1.2951      1.089      1.189      0.234      -0.839       3.430
petal_width     15.9227      3.981      4.000      0.000       8.121      23.725
================================================================================
```

How is it possible that the coefficient for sepal width is -4.8 in one model but 0.4 in the other model? Given this result, would we expect sepal width and petal width to be positively or negatively correlated?

Solution: When ignoring petal width, increasing sepal width increases the log-odds ratio of belonging to Virginica, but decreases the ratio when also accounting for petal width. This suggests that sepal and petal widths are negatively correlated.

4. Suppose you're curious about the average number of chocolate cupcakes Cupcakin' Bake Shop sells per minute on a given day. This number depends on the number of people who pass by the bakery on that day.

   For a given day i, let Y_{ij} denote the number of cupcakes sold during minute j of day i, let X_i denote the number of people who pass the bakery on day i, and let $\lambda(X_i)$ denote the average number of cupcakes sold per hour on day i.

   (a) Write down a plausible distribution for $Y_{ij}|X_i$. (*Hint:* There are multiple possible answers.)

   Solution:

$$Y_{ij} \sim \text{Poisson}(\frac{\lambda(X_i)}{60}). \tag{4}$$

   (b) Suppose you learn that $\lambda(X_i) = e^{\beta_0+\beta_1 X_i}$ for some parameters β_0, β_1. What is the link function g such that

$$\lambda(X_i) = g^{-1}(\beta_0 + \beta_1 X_i)? \tag{5}$$

   Solution:

$$g(z) = \ln z. \tag{6}$$

Feedback Form

On a scale of 1-5, where 1 = much too slow and 5 = much too fast, how was the pace of the discussion section?

1 2 3 4 5

Which problem(s) did you find most useful?

Which were least useful?

Any other feedback?