Linear regression and Bayesian models: predicting soil erosion 48 hours after a rainfall event

Cimmaron Yeoman

2023-04-17

Introduction: rainfall and soil erosion measurements (mm) after 48 hours

This data set includes 200 observations of rainfall and soil erosion measurements (mm) recorded 48 hours later. This report will assess the relationship between soil erosion and rainfall, and predict soil erosion from with a linear regression. Using this model, a prediction plot with a line of best fit and 91% prediction interval will be made. This will be repeated with a Bayesian framework for comparison.

The rainfall and soil erosion set was imported into RSTudio and renamed **rain_ero**.

Exploring the data

The average amount of rainfall was **66.8 mm** and the average soil erosion was **35.87 mm**. 75% of rainfall measurements were **103.3 mm** or less, and 75% of the soil erosion measurements were **48.0 mm** or less. This information will be used to help define a grid for prediction intervals and HPDI (highest probability density intervals).

```
##   rainfall_mm    soil_erosion_mm
##  Min.   :  1.0   Min.   : 0.00
##  1st Qu.: 29.5   1st Qu.:23.00
##  Median : 62.9   Median :35.50
##  Mean   : 66.8   Mean   :35.87
##  3rd Qu.:103.3   3rd Qu.:48.00
##  Max.   :149.4   Max.   :79.00
```

There were no signs of extreme values in the distributions of either variable.

Figure 1

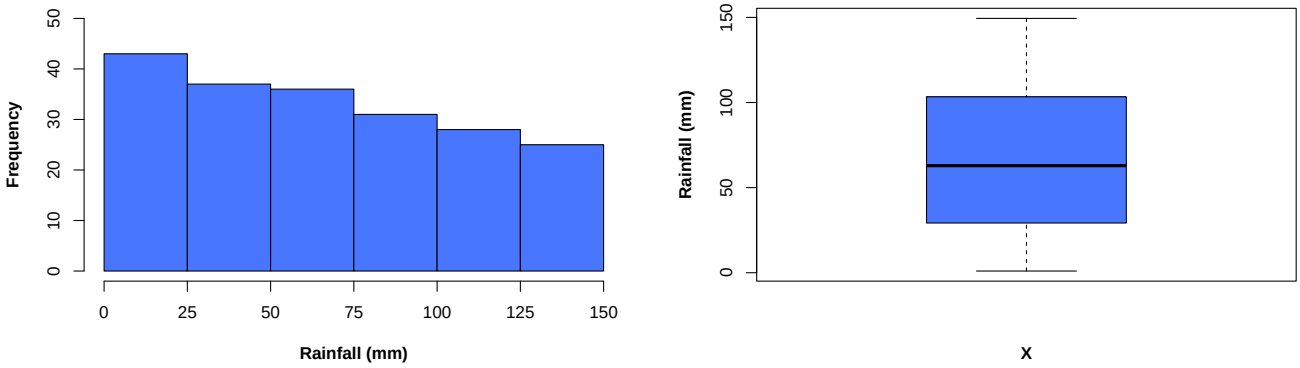
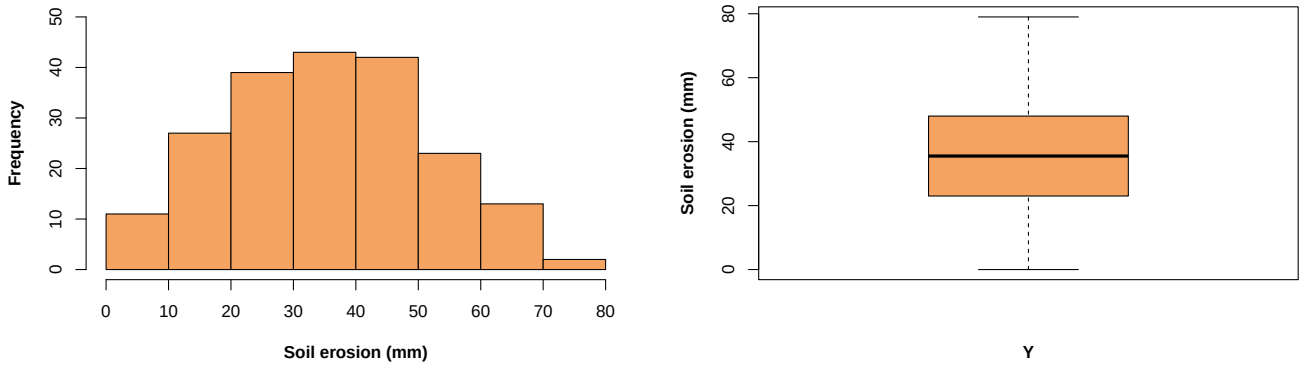


Figure 2



A simple linear model

First, we will predict the response of soil erosion in millimeters, from the amount of rainfall 48 hours prior.

```
modRE1 <- lm(soil_erosion_mm ~ rainfall_mm, data = rain_ero)
```

```
##                Estimate Std. Error  t value     Pr(>|t|)
## (Intercept) 13.5997421 1.01895350 13.34677 2.106879e-29
## rainfall_mm  0.3333946 0.01281296 26.02011 8.259021e-66
```

Model **modRE1** had an adjusted R^2 of 77.3% variation explained, indicating a strong, positive correlation. From the coefficients we can see that as soil erosion increases, rainfall also increases.

Prediction plot

Figure 3 is a prediction plot of the data set with **modRE1** as the line of best fit, and 91% HPDI and prediction intervals. The prediction interval shows the area where **modRE1** expects to find 91% of the true soil erosion values at each rainfall amount.

Figure 3

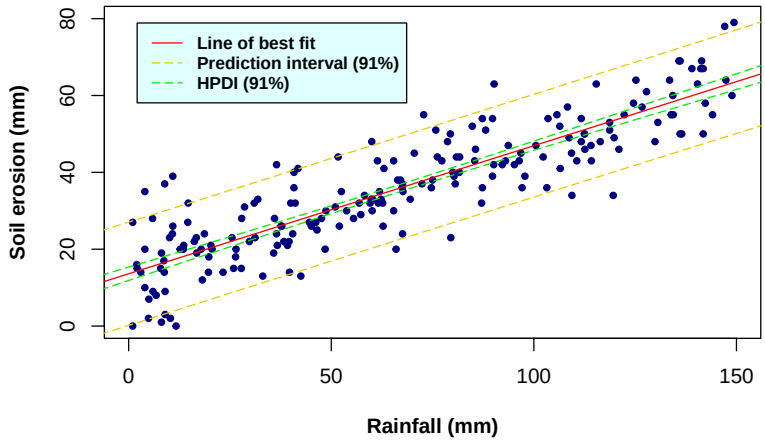


Figure 4

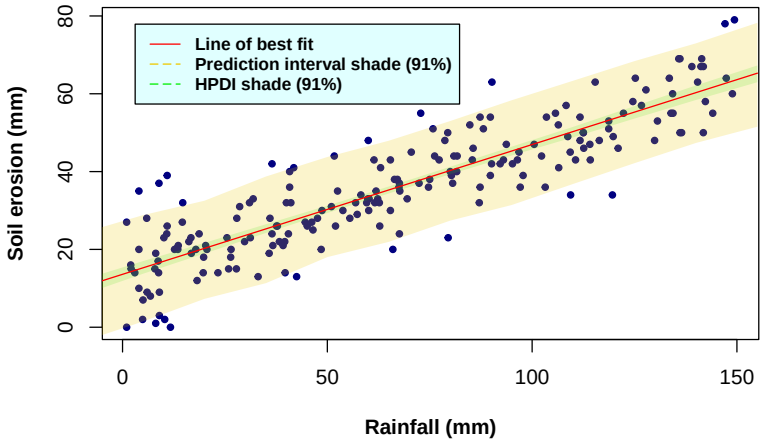


Figure 4 is a replica of Figure 3, using the shade function to indicate the prediction and HPDI areas, at 91%.

Parametric confidence intervals

These are the parametric confidence interval limits for **modRE1**. If this study were conducted and analyzed many times over, then 91% of the intervals calculated would include the true values of the parameters (our model coefficients). This is only conditional on the particular model, **modeRE1**, if I am correct. I base these assumptions off of the theory described in Statistical Rethinking, as confidence intervals are often misinterpreted (I am guilty here).

```
##                   4.5 %     95.5 %
## (Intercept) 11.8637182 15.3357660
## rainfall_mm  0.3115647  0.3552244
```

Bayesian framework

Basic, non-specific priors

For this model, I started with basic priors for the parameters to assess. This Bayesian model was almost exactly the same as the basic linear model. The intercept and slope were quite similar to **modRE1** (**0.334** versus **0.333**, **13.565** versus **13.599**). Sigma can be compared to the residual standard error of **modRE1** (**7.78** versus **7.82**).

```
modRE_B <- map(alist(
              soil_erosion_mm ~ dnorm(mu, sigma),
              mu <- a + b * rainfall_mm,
              a ~ dnorm(0, 20),
              b ~ dnorm(0, 20),
              sigma ~ dunif(0, 100)
            ),
            data = rain_ero)
```

```
##              mean         sd       4.5%      95.5%
## a     13.5653156 1.01264763 11.8484751 15.2821561
## b      0.3337575 0.01273846  0.3121607  0.3553542
## sigma  7.7810668 0.38910943  7.1213716  8.4407620
```

Specifiying values

We can modify the slope, intercept, and sigma values to try and create more informative priors. In **modRE2**, I have changed the priors to be much more specific. As an environmental science major I know a bit about dirt and rain, and I have have the knowledge that there is a positive relationship between these variables. Even though I have a basic understanding of the data, I did not have to examine samples to create more specific priors. The priors I chose below make **modRE2** even more similar to **modRE1**.

The sigma value (**7.780**) says that 91% of the possible soil erosion amounts fall within two standard deviations of the mean soil erosion amount (**35.87 mm** according to our initial summary of soil erosion). The table below also includes uncertainty of sigma, which has a 91% interval of **7.121** to **8.440**

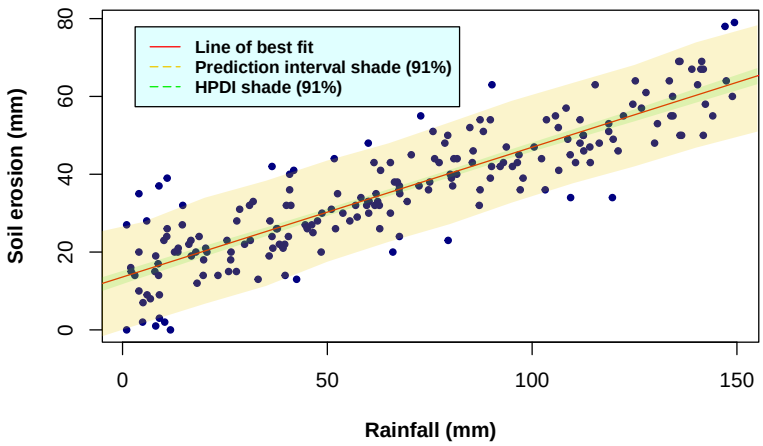
```
modRE2 <- map(alist(
              soil_erosion_mm ~ dnorm(mu, sigma),
              mu <- a + b * rainfall_mm,
              a ~ dnorm(10, 18),
              b ~ dnorm(0, 5),
              sigma ~ dunif(0,8)
            ),
            data = rain_ero)
```

```
##             mean        sd       4.5%      95.5%
## a     13.5885026 1.0122424 11.8723491 15.3046560
## b      0.3335143 0.0127345  0.3119243  0.3551043
## sigma  7.7803455 0.3890188  7.1208040  8.4398871
```

Bayesian model prediction plot

The Bayesian model prediction plot (Figure 3) is indistinguishable from the simple linear regression prediction plot, as there was very little difference between the **modRE1** and **modRE2** values. The line of best fit can be considered the average of the population mean values. The HPDI shade is sort of like a confidence interval for the expected mean values. As mentioned for Figure 3, the prediction interval captures the area on the plot where the Bayesian model (**modRE2**) expects to find 91% of the actual soil erosion values, at specific rainfall amounts.

Figure 5




```
##             mean        sd       4.5%      95.5%
```

```
## a     13.5885026 1.0122424 11.8723491 15.3046560
## b      0.3335143 0.0127345  0.3119243  0.3551043
## sigma  7.7803455 0.3890188  7.1208040  8.4398871
```

Providing less helpful values

In **modRE2**, I first experimented with different prior values that were uninformative and a bit wacky. At a certain point, the model will not run and a message will prompt you to change your values. If you change the slope and intercept, the data still overpowers the uninformative priors, and the values do not change much. Sigma seems much more sensitive to any extreme changes and produces an error code more often.

Conclusion

The choice of priors in a Bayesian linear model appear to be reasonably flexible and forgiving, at least for this report and data set. There are various acceptable priors, and many of them produce models comparable to the simple linear regression, with similar parameter values. This does not make priors useless though. Well thought priors can help us produce useful models with parameters which are not nonsensical. In this report, basic soil science and physics concepts would act as the knowledge source to create our priors. I think most people even without high level knowledge on this topic would agree that it would not make sense for a torrential downpour to cause no erosion, especially on a fluffy soils, ill-protected from the eroding actions of rain or wind. Changing the priors also lets us explore different models and how their parameters change when our initial information is different.

Both highly specific and unspecific priors still may potentially influence model parameters, even if most of the time the data or samples can overcome a wackier prior. Priors and posteriors are important to consider especially in more serious, practical applications such as a medical tests for a rare or life-threatening disease. It would be wise to consider the likelihood of a patient receiving a positive result for the disease, according to the population diagnostic rate for example. Facts making the individual an even more likely candidate for a true positive result would also be valuable (relevant symptoms most likely). Priors allow us to give our model useful information before we look at the data, allowing us to include our assumptions and knowledge, without secretly observing a sample and deciding on priors to produce a certain outcome (this is not a good idea!).