Comparison of S1 protein conditions

Hefin Rhys

14 November, 2020

Purpose

This is an R Markdown document detailing my analysis of Peter's ELISA data. The purpose of the experiment is to determine whether different reactivity is observed between five different SARS-Cov2 S1 spike protein conditions. 90 samples (with varying degrees of reactivity) were each tested across the five conditions, and the data therefore represent repeated measurements made on each sample.

Loading packages

I start by loading the required R packages for the analysis.

```
library(tidyverse)                              # data wrangling and plotting
library(rethinking)                             # Bayesian modelling
library(tidybayes)                              # tools for tidying model draws
library(tidybayes.rethinking)                   # tools for tidying model draws
library(modelr)                                 # for data_grid() function
```

Reading data

Next, I read in the data that is in tidy format as a .csv file, and store it as the object `elisa`.

```
elisa <- read_csv("../data/ELISA_data.csv")
```

I inspect its structure by printing the first 10 rows of data. There are four columns: a character vector `Protein` that indicates the experimental condition, a double precision numeric vector `OD` that indicates the optical density of the reading (the dependent variable), a double precision numeric vector `SCO` that will not be used in this analysis, and a character vector `Sample`, which indicates which sample the reading was taken from.

```
elisa
```

```
## # A tibble: 450 x 4
##    Protein    OD    SCO Sample
##    <chr>   <dbl>  <dbl> <chr>
##  1 S1       0.97   5.94 CovIC106
##  2 S1       0.83   5.08 CovIC111
##  3 S1       2.69  16.4  CovIC113
##  4 S1       0.65   3.98 CovIC105
##  5 S1       0.3    1.81 CovIC108
##  6 S1       0.33  2.03  CovIC112
##  7 S1       0.11  0.656 CovIC114
##  8 S1       0.05  0.292 CovIC107
##  9 S1       0.05  0.307 CovIC110
## 10 S1       0.39  2.40  CovIC100
## # ... with 440 more rows
```

Plotting the empirical data

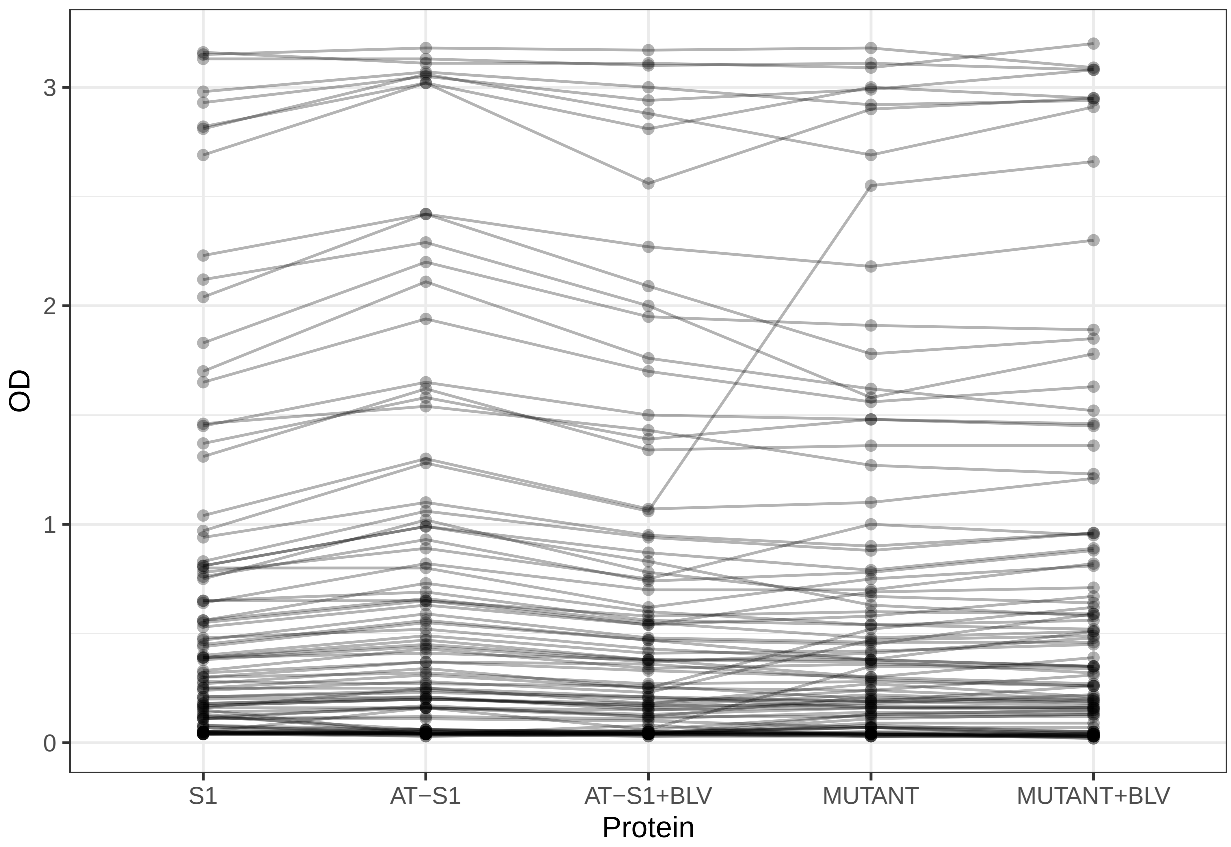
I next plot the data to visualize the relationships between the `Protein`, `Sample`, and `OD` variables. So that the levels of the `Protein` variable are plotted in order, I first convert this variable into a factor and specify the order of its levels.

```
elisa$Protein <- factor(
  elisa$Protein,
  levels = c("S1", "AT-S1", "AT-S1+BLV", "MUTANT", "MUTANT+BLV")
  )
```

Now that `Protein` is a factor, its levels will be plotted in the desired order. There are a few observations to make about this data:

- the dependent variable is strictly positive and is clearly bounded at zero, but does not have an obvious upper bound
- the marginal distribution of the dependent variable is clearly not normally-distributed, but is considerably positively skewed
- OD values between conditions are highly correlated within samples, and ignoring the repeated observations will likely underestimate parameter estimates

```
ggplot(elisa, aes(x = Protein, y = OD, group = Sample)) +
  geom_line(alpha = 0.3) +
  geom_point(alpha = 0.3) +
  theme_bw()
```



Modeling the data

To model the relationships in the data, I start by numerically-encoding the categorical variables, and removing the `SCO` variable. This is just how the `ulam()` function I use to fit the models, expects categorical variables.

```
model_data <- elisa %>%
  mutate(Protein = as.integer(Protein),
         Sample = as.integer(as.factor(Sample))) %>%
  select(-SCO)
```

```
$$\textrm{OD} \sim \textrm{Gamma}(\mu, \textrm{scale}),$$
$$\textrm{log(}\mu) = \textrm{intercept}[\textrm{Sample}] + \textrm{offset}[\textrm{Protein}],$$
$$\textrm{offset}[\textrm{Protein}] \sim  \textrm{Normal}(0, 1),$$
$$\textrm{scale} \sim \textrm{Exponential}(5),$$
$$\textrm{intercept}[\textrm{Sample}] \sim \textrm{Gamma}(\mu_\textrm{intercept}, \textrm{scale}_\te
$$\mu_\textrm{intercept} \sim \textrm{Normal}(1, 0.2),$$
$$\textrm{scale}_\textrm{intercept} \sim \textrm{Exponential}(5)$$
```