Test 2 – Take Home Problems

Stat 341 — Spring 2017

Loading some data

The command below will load some objects for this test. In particular, this will load

- the model objects `u1`, . . . , `u7` (so you don't have to wait for Stan to fit them)
- the `Pallets2` data frame

```
# load some data for this test.
load(file = url("https://rpruim.github.io/s341/data/Test2-S21.Rda"))
```

Pallet repair

A local company repairs pallets like the ones in the image above. They are interested to know which of their four employees are more or less efficient at this. Since they don't know much statistics, they enlist your help as a consultant to answer their question.

You don't know much about pallets, so you ask them "How many pallets can a person repair in one day?" The answer: "Oh that depends on the person and on the day. I'm guessing it's usually somewhere between 100 and 150."

We'll use that information below to specify some reasonable priors.

The data they provide (in `Pallets2`) gives the number of pallets repaired by each of four workers on each of five days. I've already added indexed versions of `employee` and `day` to the data.

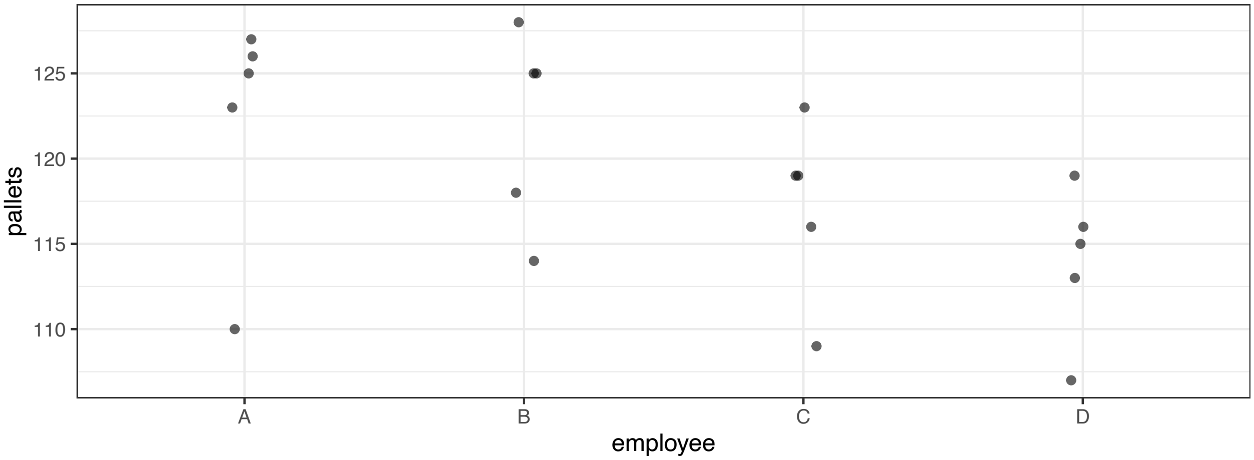
```
Pallets2 <-
  fastR2::Pallets %>%
  mutate(
    emp_idx = as.numeric(factor(employee)),
    day_idx = as.numeric(factor(day))
  )
```

```
Pallets2 %>% head(4)
```

```
##   pallets employee  day emp_idx day_idx
## 1     123        A day1       1       1
## 2     127        A day2       1       2
## 3     126        A day3       1       3
## 4     125        A day4       1       4
```

Here's a quick look at the data.

```
gf_jitter(pallets ~ employee, data = Pallets2, height = 0, width = 0.05, alpha = 0.6)
```



Two simple models

As you have seen in class, it is often good to start with a simple model and build up from there. Here are two models. (Remember, you don't need to run this code if you load the models above.)

```
set.seed(111)
u1 <- ulam(
  data = Pallets2,
  alist(
    pallets ~ dnorm(mu, sigma),
    mu <- a,
    a ~ dnorm(125, 20),
    sigma ~ dexp(1)
  ),
  chains = 4, iter = 2000, cores = 4, refresh = 0,
  file = "test2-u1"
)
```

```
set.seed(222)
u2 <- ulam(
  data = Pallets2,
  alist(
    pallets ~ dnorm(mu, sigma),
    mu <- b[emp_idx],
    b[emp_idx] ~ dnorm(125, 20),
    sigma ~ dexp(1)
  ),
  chains = 4, iter = 2000, cores = 4, refresh = 0,
  file = "test2-u2"
)
```

1. Which model fits the data (in `Pallets2`) better? How do you know?

2. Which employee does model `u2` think is best (most efficient)? Worst?

3. How sure is model `u2` that the most efficient employee is more efficient than the least efficient employee? Use posterior sampling to quantify your answer.

Day to day

Time for another model. This time we are looking at how the number of pallets repaired by one employee depends on the day.

```
set.seed(333)
u3 <- ulam(
  data = Pallets2,
  alist(
    pallets ~ dnorm(mu, sigma),
    mu <- d[day_idx],
    d[day_idx] ~ dnorm(125, 20),
    sigma ~ dexp(1)
  ),
  chains = 4, iter = 2000, cores = 4, refresh = 0,
  file = "test2-u3"
)
```

4. What does u3 say about day to day variation in the number of pallets repaired?

5. Comparing u2 and u3, does one of employee or day appear to have a greater impact on the number of pallets repaired? Explain.

Employees and Days

Below are four models that include both day and employee.

```
set.seed(444)
u4 <- ulam(
  data = Pallets2,
  alist(
    pallets ~ dnorm(mu, sigma),
    mu <- b[emp_idx] + d * day_idx,
    b[emp_idx] ~ dnorm(125, 20),
    d ~ dnorm(0, 10),
    sigma ~ dexp(1)
  ),
  chains = 4, iter = 2000, cores = 4, refresh = 0,
  file = "test2-u4"
)
```

```
set.seed(555)
u5 <- ulam(
  data = Pallets2 %>% mutate(day_idx2 = day_idx^2),
  alist(
    pallets ~ dnorm(mu, sigma),
    mu <- b[emp_idx] + d1 * day_idx + d2 * day_idx2,
    b[emp_idx] ~ dnorm(125, 20),
    c(d1, d2) ~ dnorm(0, 10),
    sigma ~ dexp(1)
  ),
  chains = 4, iter = 4000, warmup = 1000, cores = 4, refresh = 0,
  file = "test2-u5"
)
```

```
set.seed(666)
u6 <- ulam(
  data = Pallets2,
  alist(
    pallets ~ dnorm(mu, sigma),
    mu <- b[emp_idx] + d[day_idx],
    b[emp_idx] ~ dnorm(125, 20),
    d[day_idx] ~ dnorm(0, 10),
    sigma ~ dexp(1)
  ),
  chains = 4, iter = 4000, warmup = 1000, cores = 4, refresh = 0,
  file = "test2-u6")
```

```
set.seed(777)
u7 <- ulam(
  data = Pallets2,
  alist(
    pallets ~ dnorm(mu, sigma),
    mu <- b[emp_idx] + d[day_idx],
    b[emp_idx] ~ dnorm(0, 10),
    d[day_idx] ~ dnorm(125, 20),
    sigma ~ dexp(1)
  ),
  chains = 4, iter = 4000, warmup = 1000, cores = 4, refresh = 0,
  file = "test2-u7"
)
```

6. The only difference between the definitions `u6` and `u7` is the priors. What does this difference in priors do?

7. What would happen if we used `dnorm(125, 20)` for both (sets of) normal priors? Why is this not as good as `u6` and `u7`?

8. What would happen if we used `dnorm(0, 10)` for both (sets of) normal priors? Why is this not as good as `u6` and `u7`?

9. For each of these four models, how many pallets does the model predict employee #2 would repair on day 2?

Choose Wisely

10. Time to pick your favorite model from among `u4`, `u5`, `u6`, and `u7` (for the purpose of determining whether the employees differ). Explain how you chose it and provide at least one plot to show how well/poorly it works.

11. Are there any indications of problems with the `ulam()` fit for your favorite model? Make it clear what things you have looked at and how you decided whether they look good or not. (This should probably have been asked earlier, but I didn't want you to have to do it for all of the models.)

12. How sure is your chosen model that the best (most efficient) employee is better than the worst? Quantify your answer.

13. How does your answer in 12 compare to your answer in 3? Why?