19-8-9

Nikhil Gopal

3/29/2022

19.8-9

Messy randomization: The folder Cows contains data from an agricultural experiment that was conducted on 50 cows to estimate the effect of a feed additive on 6 outcomes related to the amount of milk fat produced by each cow

Four diets (treatments) were considered, corresponding to different levels of the additive, and three variables were recorded before treatment assignment: lactation number (seasons of lactation), age, and initial weight of the cow.

Cows were initially assigned to treatments completely at random, and then the distributions of the three covariates were checked for balance across the treatment groups; several randomizations were tried, and the one that produced the "best" balance with respect to the three covariates was chosen. The treatment depends only on fully observed covariates and not on unrecorded variables such as the physical appearances of the cows or the times at which the cows entered the study, because the decisions of whether to re-randomize are not explained. We shall consider different estimates of the effect of additive on the mean daily milk fat produced.

- a) Consider the simple regression of mean daily milk fat on the level of additive. Compute the estimated treatment effect and standard error, and explain why this is not a completely appropriate analysis given the randomization used.

```
data <- read.csv("cows.csv")

fit <- stan_glm(fat~level, data = data, refresh = 0)

print(fit)
```

```
## stan_glm
##  family:       gaussian [identity]
##  formula:      fat ~ level
##  observations: 50
##  predictors:   2
## ------
##             Median MAD_SD
## (Intercept) 3.3    0.1
## level       2.0    0.6
##
## Auxiliary parameter(s):
##       Median MAD_SD
## sigma 0.4    0.0
##
```

```
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

The estimated treatment effect was 2.0 with an SE of 0.5. However, the two groups were selected in order to best account for a desired distribution of covariates. It is necessary to assign treatments first, and then record covariate measures afterwards.

- b) Add more predictors to the model. Explain your choice of which variables to include. Compare your estimated treatment effect to the result from (a).

```
data$lactation <-as.factor(data$lactation)

fit_2 <- stan_glm(fat~level+age+lactation+initial.weight, data = data, refresh = 0)

print(fit_2)
```

```
## stan_glm
##  family:       gaussian [identity]
##  formula:      fat ~ level + age + lactation + initial.weight
##  observations: 50
##  predictors:   9
## ------
##                Median MAD_SD
## (Intercept)    2.8    0.7
## level          2.0    0.6
## age            0.0    0.0
## lactation2     0.2    0.2
## lactation3     0.5    0.4
## lactation4     0.5    0.6
## lactation5     0.9    0.7
## lactation6     1.7    0.9
## initial.weight 0.0    0.0
##
## Auxiliary parameter(s):
##       Median MAD_SD
## sigma 0.4    0.0
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

In this model, the treatment effect on the level variable was 2.0 which was the same, and the standard error was 0.6, which was almost the exact same.

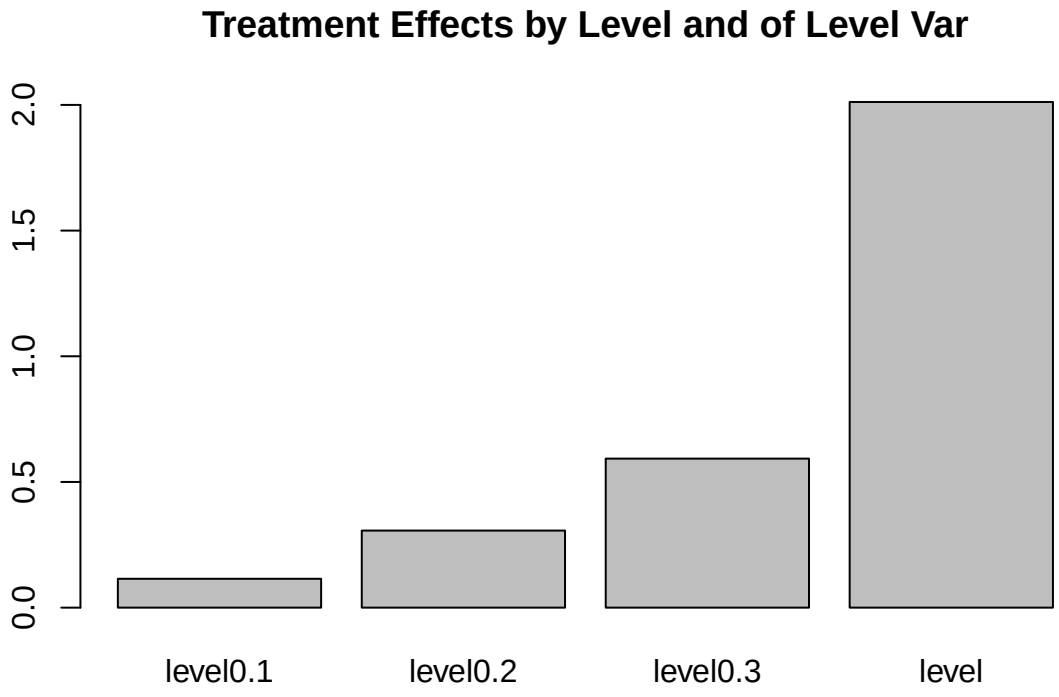
I added the age variable, the lactation variable, and the initial weight variable. I figured that older cows would be likely to produce less milk, and that the lactation variable would be associated with if a specific cow was likely to produce a lot or a little milk. Initial weight was added because some deformed cows might affect the data set.

- c) Repeat (b), this time considering additive level as a categorical predictor with four levels.

```
data$level <- as.factor(data$level)
fit_3 <- stan_glm(fat~level+age+lactation+initial.weight, data = data, refresh = 0)
```

Make a plot showing the estimate (and standard error) of the treatment effect at each level, and also showing the inference from the model fit in part (b).

```
treatment_effects <- c(coef(fit_3)[2:4],coef(fit_2)[2])
barplot(treatment_effects, main = "Treatment Effects by Level and of Level Var")
```



Causal inference based on data from individual choices: Our lives involve trade offs between monetary cost and physical risk, in decisions ranging from how large a car to drive, to choices of health care, to purchases of safety equipment. Economists have estimated how people implicitly trade off dollars and danger by comparing choices of jobs that are similar in many ways but have different risks and salaries. This can be approximated by fitting regression models predicting salary given the probability of death on the job (and other characteristics of the job). The idea is that a riskier job should be compensated with a higher salary, with the slope of the regression line corresponding to the "value of a statistical life."

- a) Set up this problem as an individual choice model, as in Section 15.7. What are an individual's options, value function, and parameters?

We will built a logistic regression model to predict the probability of one switching from a less risky to a risky job as a function of remuneration received. Jobs will be classified as risky or less risky based on their workplace death/incident rates. The parameters will be the worker's salary, their already accumulated wealth, the amount of money they need to sustain their family, their partner's salary (0 if no partner) and their education level.

The value function will essentially be the model generated, and will predict at what income level an individual might choose to switch between risky/unrisky jobs, with all other variables held constant.

- b) Discuss the assumptions involved in assigning a causal interpretation to these regression models.

This model assumes that individuals make choices based on the parameters given, which is obviously a false asssumption. There are many confounders that are not present in my model. Another assumption is that parameters do not change over time. Variables that I have included like wealth and spouses wealth or education do indeed change over time. A choice model might be a useful tool in situations like this, but it definitely does rely on assumptions that are fundamentally impossible to meet.