STA 610 LAB 3

William Tirone

2023-09-06

EDA

First, we notice that there are some null values that present issues with fitting the model, so we remove them. In adddition, after plotting a density of the net income, we see a very large range of values, so I chose to scale them. We can't take the log because there are some negative net income values.

```
hc2014 = hc2014 |>
  drop_na() |>
  mutate(scaled_income = scale(netincome))

agg = hc2014 |>
  group_by(control) |>
  summarise(total_beds = sum(numbeds)) |>
  arrange(desc(total_beds))

agg[0:10,]
```

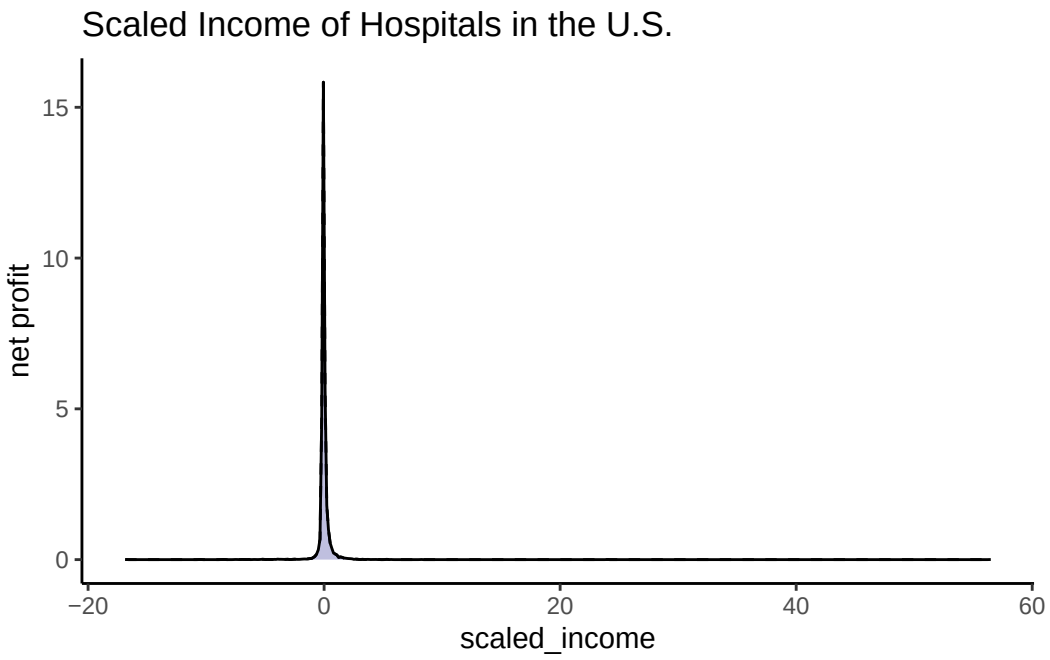
```
# A tibble: 10 x 2
   control                total_beds
   <fct>                       <dbl>
 1 Nonprofit-other            383324
 2 Corporation                166482
 3 Nonprofit-church           104345
 4 Gvmt-State                  46498
 5 Gvmt-County                 31339
 6 Gvmt-Hospital District      27978
 7 Gvmt-Other                  12480
 8 Partnership                 10317
```

```
 9 Gvmt-City-County              9225
10 Gvmt-City                     7935
```

We see a very sharp spike at 0, but also some pretty large outliers.

```
ggplot(hc2014, aes(scaled_income)) +
  geom_density(aes(y=..density..),color="black",linetype="dashed") +
  theme(legend.position="none") +
  geom_density(alpha=.25, fill="navy") +
  labs(title="Scaled Income of Hospitals in the U.S.",y="net profit") +
  theme_classic()
```

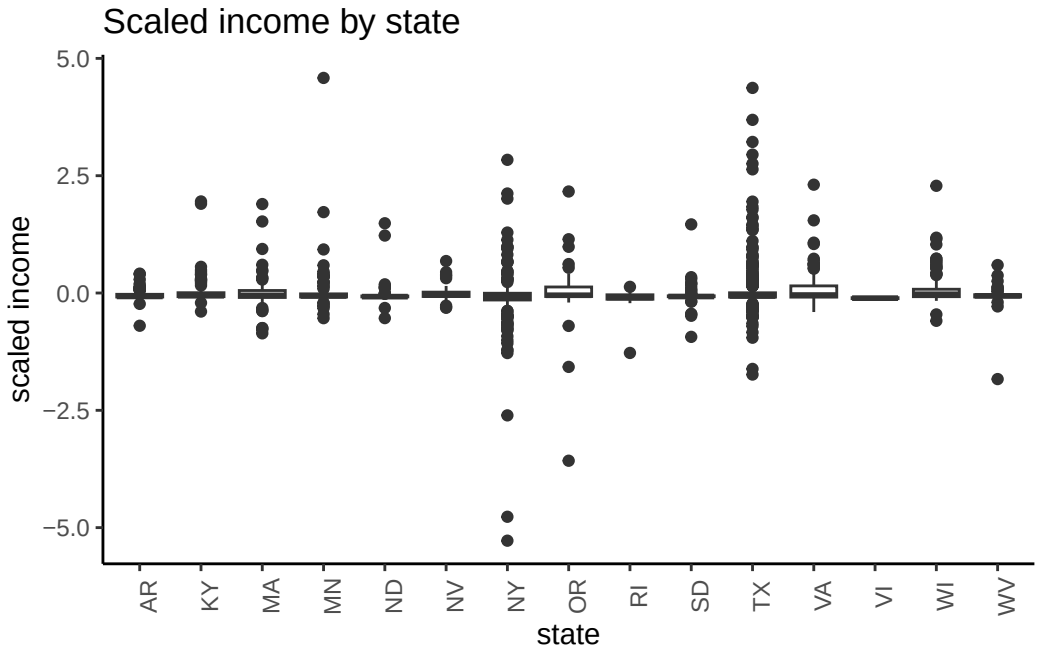
```
Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
i Please use `after_stat(density)` instead.
```



Below, it looks like with a random sample of states, we have a pretty significant spread among them. New York has some hospitals with pretty large negative scaled profits and Texas has some with very positive values.

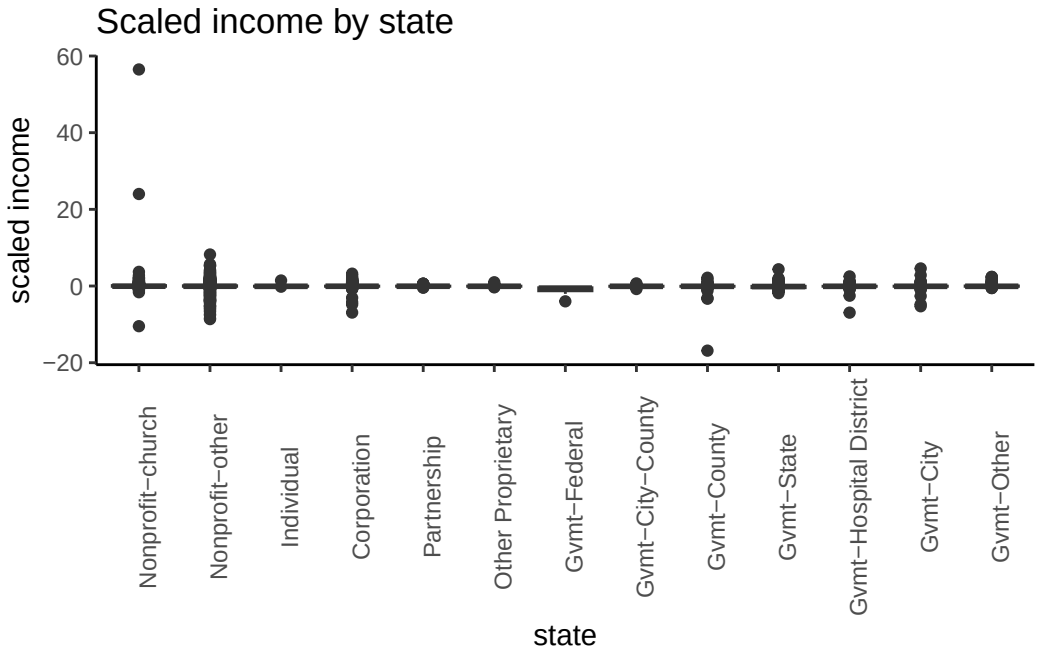
```
set.seed(1000)
sample_state <- sample(unique(hc2014$state),15,replace=F)
```

```
ggplot(hc2014[is.element(hc2014$state, sample_state),],
       aes(x=state, y=scaled_income)) +
  geom_boxplot() +
  labs(title="Scaled income by state",
       x="state",y="scaled income") + theme_classic() +
  theme(legend.position="none",axis.text.x = element_text(angle = 90))
```



And considering control, there's large variability in the `Nonprofit-church` ownership. It's interesting that `Nonprofit-other` has the largest number of beds by far, but they have much less variability.

```
ggplot(hc2014,
       aes(x=control, y=scaled_income)) +
  geom_boxplot() +
  labs(title="Scaled income by state",
       x="state",y="scaled income") + theme_classic() +
  theme(legend.position="none",axis.text.x = element_text(angle = 90))
```



Modeling

Our final model will be :

$$y_{ij} = \mu_0 + \alpha_j + \beta_1 * \text{num beds} + \beta_2 * \text{control} + \epsilon_{ij}$$

We will fit 2 models and compare them, a standard linear regression and a model with a random effect.

Neither of the random slope models we tried to fit in lab converged or would run. In particular:

```
model3 = lmer(netincome ~ numbeds +  control + (1 + control | state), data=hc2014)
```

```
model3 = lmer(netincome ~ numbeds +  control + (numbeds + control | state),
data=hc2014)
```

```
model1 = lm(scaled_income ~ numbeds + control, data=hc2014) # standard
model2 = lmer(netincome ~ numbeds + control +  (1 |state), data=hc2014) # random effect
lrtest(model2, model1)
```

```
Warning in modelUpdate(objects[[i - 1]], objects[[i]]): original model was of
class "lmerMod", updated model is of class "lm"
```

```
Likelihood ratio test

Model 1: netincome ~ numbeds + control + (1 | state)
Model 2: scaled_income ~ numbeds + control
  #Df  LogLik Df  Chisq Pr(>Chisq)
1  16 -121645
2  15   -8615 -1 226060  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Because we observe a p-value < 0.05, we will reject the null hypothesis that both models fit the data equally well. Thus, we conclude that we want to keep the model with the random effect rather than the simple linear model.

Now, we can look at the fixed effects coefficients with their respective standard errors:

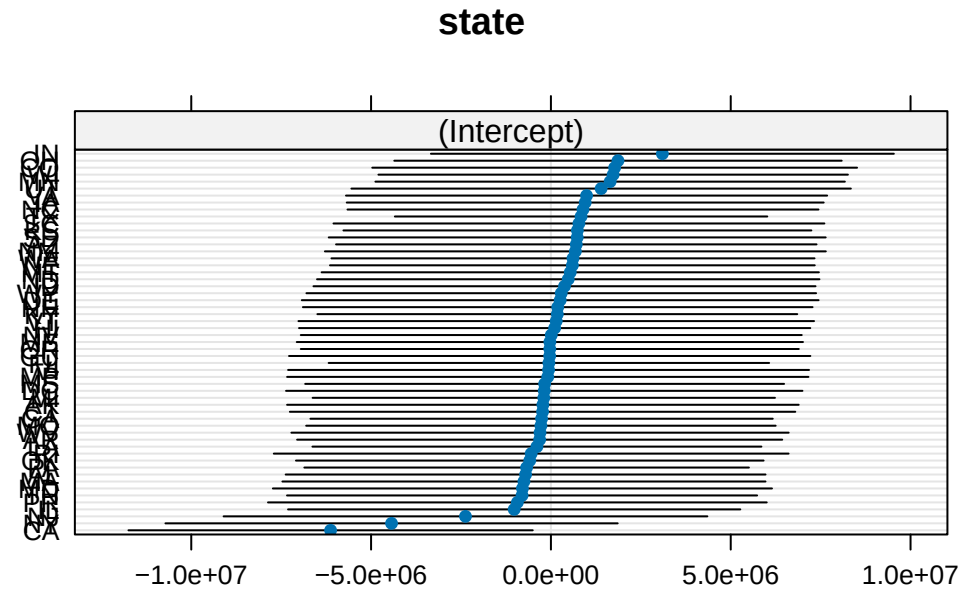
```
coef(summary(model2))
```

```
                                Estimate   Std. Error    t value
(Intercept)                    7259876.9  4003254.696  1.8134936
numbeds                         101315.4     7044.451 14.3822952
controlNonprofit-other        -13184254.7  4214050.326 -3.1286420
controlIndividual             -10538126.1 24209263.186 -0.4352931
controlCorporation            -11435963.7  4496205.262 -2.5434701
controlPartnership             -6748597.7  7848296.520 -0.8598806
controlOther Proprietary       -8348467.1  9678452.271 -0.8625829
controlGvmt-Federal           -86469599.4 14860200.757 -5.8188715
controlGvmt-City-County       -14558699.7 10110371.438 -1.4399767
controlGvmt-County            -16813418.3  5935319.241 -2.8327741
controlGvmt-State             -37104410.6  7075197.400 -5.2442933
controlGvmt-Hospital District -13399278.6  6249812.076 -2.1439490
controlGvmt-City              -18180789.2 11714552.152 -1.5519833
controlGvmt-Other              -1627110.7 11497265.451 -0.1415215
```

And we can view the random intercept and each intercept's respective uncertainty (that is, if I'm understanding this plot correctly).

```
dotplot(ranef(model2, condVar=TRUE))
```

```
$state
```



Model Assessment

As a quick check, our residuals don't show any obvious patterns.

```
plot(model2)
```