Engage3_Report

January 11, 2020

Blake Shurtz, Job Candidate

This report proceeds in two parts. First, the data join and cross tabulation. See the attached .csv. Then, the exploratory analysis/model build. In short, there is definitely an issue with the quality of the data, but after removing the bad data, there are no suprises with the results of the model.

[1]:
```
import numpy as np
```

[2]:
```
import pandas as pd
```

[3]:
```
auditors = pd.read_csv('data/auditors.csv');
```

[4]:
```
prices = pd.read_csv('data/prices.csv');
```

[5]:
```
stores = pd.read_json('data/stores.json')
```

Join Data

[6]:
```
df = pd.merge(prices, stores, how='left', on='Store ID')
```

[7]:
```
df = pd.merge(df, auditors, how='left', on='Auditor ID')
```

[8]:
```
df = df.drop(['Region_y'], axis = 1)
```

[9]:
```
df.columns = ['Auditor ID', 'Date', 'Price', 'Store ID', 'UPC', 'Banner',␣
↪'Region', 'First', 'Last']
```

[10]:
```
table = df.pivot_table(index = ['Banner', 'UPC'], columns='Region',␣
↪values='Price')
```

[11]:
```
table.head()
```

[11]:
```
Region            Kansas  New York  Northern California  Texas
Banner  UPC
Safeway 11873171     NaN       NaN                 6.09   5.19
        15052612   53.99       NaN                  NaN  54.49
        16482322   17.89       NaN                  NaN  18.09
        16729338    7.99       NaN                 9.39   8.09
        16829288    3.59       NaN                 4.19   3.59
```

[12]:
```
table.to_csv('data/table.csv')
```

Exploratory Analysis First of all, the Whole Foods in Kansas has a bunch of items that are $1.99. I understand that this could be some kind of loss-leader situation, but this is unlikely.

[13]:
```
table.tail()
```

[13]:
```
Region                   Kansas  New York  Northern California  Texas
Banner      UPC
Whole Foods 995798889      1.99     62.39                70.39  60.59
            996262978      1.99     14.39                16.19  13.89
            996849471      1.99     12.79                  NaN    NaN
            998831540      1.99     39.99                  NaN  38.79
            999185078      1.99     58.09                65.49    NaN
```

[14]:
```
filtered_df = df[(df['Banner'] == 'Whole Foods') & (df['Region'] == 'Kansas')]
```

[15]:
```
filtered_df['Auditor ID'].value_counts() #frequency table
```

[15]:
```
713    913
Name: Auditor ID, dtype: int64
```

[16]:
```
filtered_df['Store ID'].value_counts() #frequency table
```

[16]:
```
39287    913
Name: Store ID, dtype: int64
```

It's either the auditor or the store. Either way, this data is no good. The data from this particular store will be removed from the model.

0.0.1 Running Model

I'll run a two-factor multi-level model using Stan. I'm excluding the UPC from the model, so that the linear model I'm using is more like P_ijk = Aj + Bk.

[17]:
```
import pystan as pystan
```

[18]:
```
stan_data = pystan.read_rdump('data/stan_data.r') #data needed to be converted␣
↪to a stan data object in R, see issue here: https://github.com/stan-dev/
↪pystan/issues/437
```

[19]:
```
my_model_code = """
data{
     int<lower=0> N_obs;
         int<lower=1> N_regions;
     int<lower=1> N_banners;
         int banner_id[N_obs];
         int region_id[N_obs];
     real Price[N_obs];
 }

parameters{
     vector[N_regions] a_r;
     vector[N_banners] a_s;
         real a;
```

```
        real<lower=0> sigma_r;
        real<lower=0> sigma_s;
    real<lower=0> sigma;
}
model{
    vector[N_obs] mu;
    sigma ~ cauchy(0,1);
    a_r ~ normal(0,sigma_r);
        sigma_r ~ cauchy(0,1);
        a_s ~ normal(0, sigma_s);
        sigma_r ~ cauchy(0,1);
    mu = a + a_r[region_id] + a_s[banner_id];
    Price ~ normal(mu, sigma);
        }
"""
```

[20]:
```
sm = pystan.StanModel(model_code=my_model_code)
```

```
INFO:pystan:COMPILING THE C++ CODE FOR MODEL
anon_model_f173c5b356c0123acd48aa644ff2a541 NOW.
```

[21]:
```
fit = sm.sampling(data=stan_data, iter=1000, chains=4)
```

Banner(a_s): Walmart = 1; Whole Foods = 2; Wegmans = 3; Trader Joes = 4; Safeway = 5
Region(a_r): Texas = 1; Kansas = 2; NY = 3; Northern CA = 4

[22]:
```
print(fit)
```

```
Inference for Stan model: anon_model_f173c5b356c0123acd48aa644ff2a541.
4 chains, each with iter=1000; warmup=500; thin=1;
post-warmup draws per chain=500, total post-warmup draws=2000.

          mean se_mean     sd   2.5%    25%    50%    75%  97.5%  n_eff   Rhat
a_r[1]   -1.18    0.04   0.94  -3.11  -1.74  -1.16   -0.6   0.72    653    1.0
a_r[2]   -1.06    0.04   0.95  -2.95  -1.63  -1.04  -0.49   0.82    627    1.0
a_r[3]   -0.53    0.04   0.95  -2.46  -1.12  -0.51   0.05   1.39    674    1.0
a_r[4]    2.92    0.04   0.95    1.1   2.31    2.9   3.49   4.91    641    1.0
a_s[1]   -2.15    0.09   1.44  -4.97  -2.89   -2.2  -1.49   1.22    280   1.01
a_s[2]    3.16    0.09   1.45   0.39   2.39   3.09    3.8   6.56    283   1.01
a_s[3]    0.03    0.09   1.45  -2.68  -0.71  -0.05   0.68   3.51    282   1.01
a_s[4]   -0.56    0.09   1.44  -3.38  -1.33  -0.63   0.07   2.87    287   1.01
a_s[5] -2.6e-3    0.09   1.42  -2.78  -0.74  -0.09   0.64   3.41    276   1.01
a        31.32    0.09   1.66  27.81  30.41  31.39  32.28  34.52    313   1.01
sigma_r   1.76    0.02   0.66   0.93   1.32   1.62   2.02   3.51   1052    1.0
sigma_s   2.94    0.07   1.72   1.23   1.89   2.49   3.43   7.53    576    1.0
sigma    17.27  2.5e-3   0.11  17.05   17.2  17.27  17.35  17.49   1990    1.0
lp__    -3.8e4     0.1   2.59 -3.8e4 -3.8e4 -3.8e4 -3.8e4 -3.8e4    664    1.0

Samples were drawn using NUTS at Sat Jan 11 01:25:19 2020.
```

```
For each parameter, n_eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor on split chains (at
convergence, Rhat=1).
```

I'm not too surprised at the results. Looking at the posterior means...

For the regions (a_r), Northern CA(4) is the most expensive, New York(3) is relatively cheap (this is a surprise), and Texas and Kansas are the cheapest (no surprise).

For the banners (a_s), Walmart(1) is the cheapest, Whole Foods(2) in the most expensive, Wegmans, TJ's and Safeway are the least pricey. I'm a bit surprised TJ's is less expensive than Safeway.

The variance around the banners is large enough that we should take the differences with a grain of salt. Still, the gap between Walmart and Whole Foods is large.

The variance around the regions is large enough that CA really stands out among the other 3.

The model converged well so the estimates are accurate within the model framework.