What Predicts Family Spending on Groceries?

Will Doyle

Update: 4/16/2019

The Problem

Understanding which families will spend money on groceries is a key analytic problem for our organization. In particular, we need to make sure that our advertising on every medium is targeted towards those that will be spending at least $1,000 per quarter on food groceries, with more advertising focused on those that spend more.

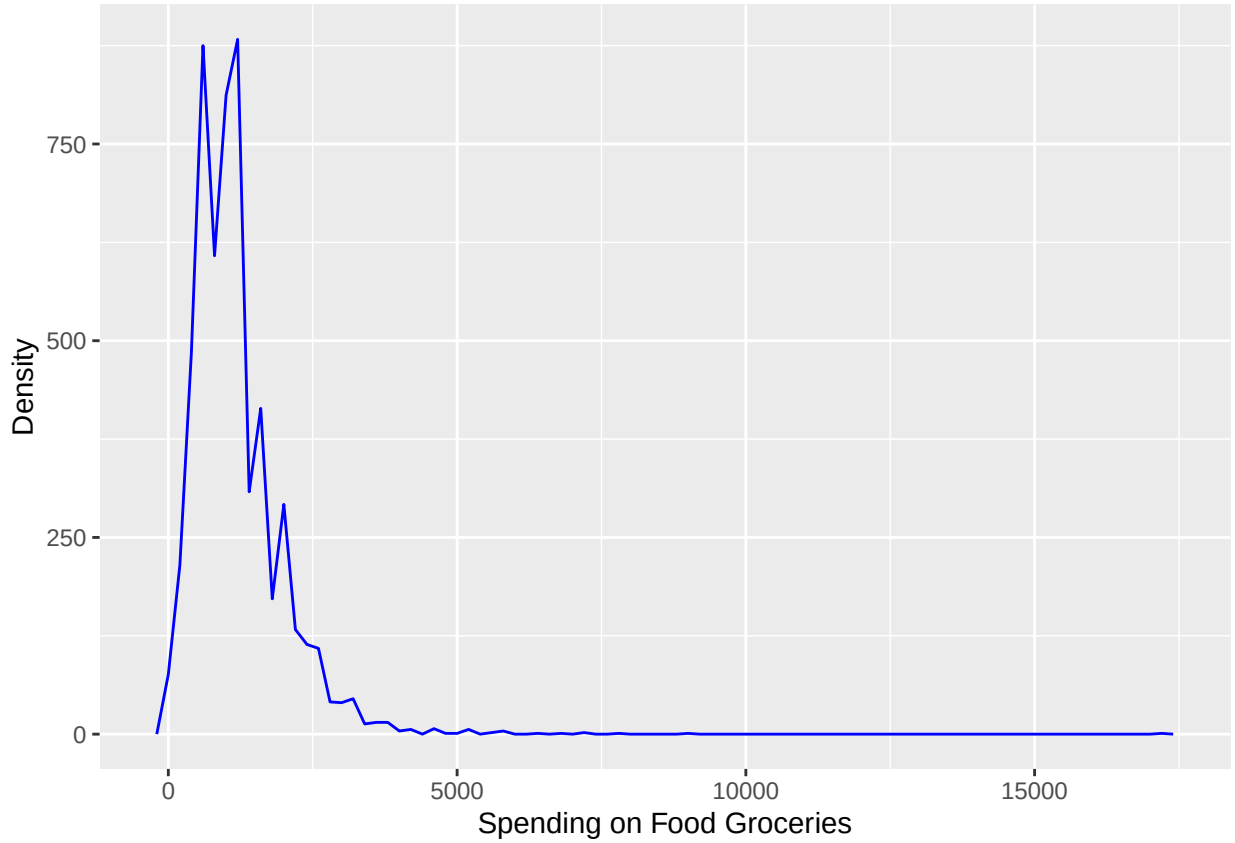
The Data

The data for this analysis come from the Consumer Expenditure Survey, a quarterly survey of a nationally represntative sample of American Families.

The figure below shows the distribution of family spending on groceries.

Spending on Groceries

```
gg<-ggplot(data=cex,aes(x=grocery_food,color=grocery_food))
gg<-gg+geom_freqpoly(binwidth=200,color="blue")
gg<-gg+xlab("Spending on Food Groceries")+ylab("Density")
gg
```




```
## What's the median spending?
groc_med<-cex%>%summarize(median(grocery_food))

groc_75<-cex%>%summarize(quantile(grocery_food,probs=.75))
```

As the figure shows, 50 percent of families spend more than 1,040 dollars, while 25 percent of families spend more than 1,560. The question for us is: what characteristics of families predict higher or lower levels of spending?

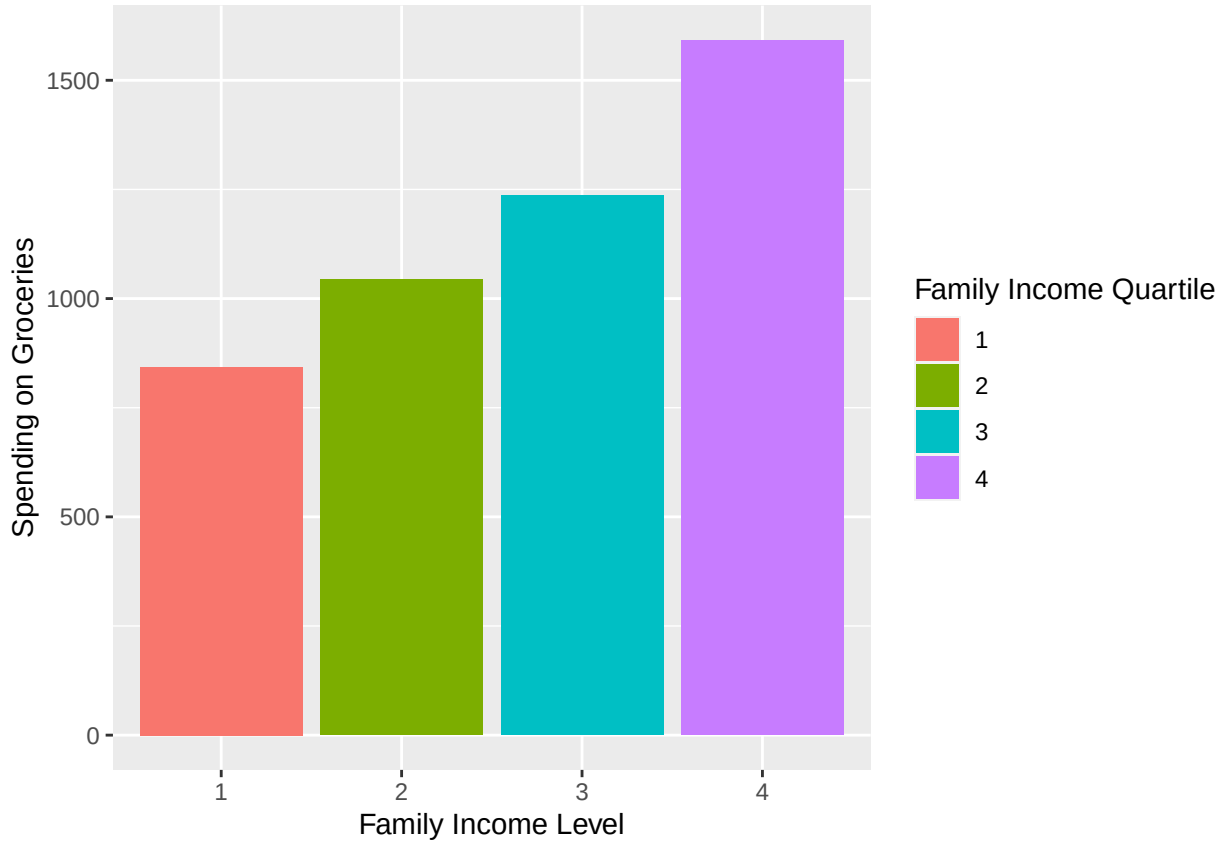
Which families spend more or less on groceries?

Families with higher income should spend more on groceries. The figure below shows average spending on groceries by family income quartile. As expected, higher income families spend more on groceries.

Spending on Groceries by Family Income Quartile

```
## By family income rank

myprob=.25 ## This gives the number of groups--.25=4 groups

## Create a variable for quantiles of college education
cex<-cex%>%mutate(fam_income_level=ntile(inc_rank,4))
```

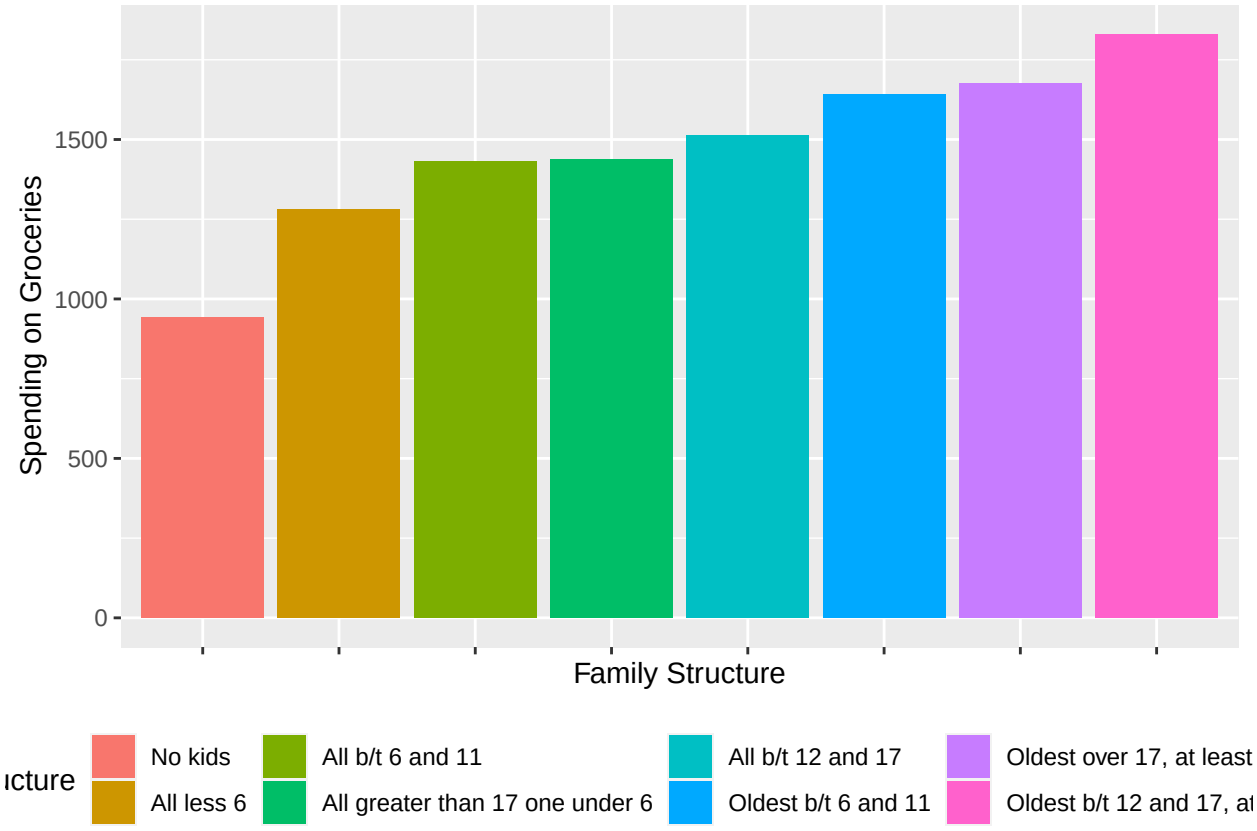
```r
cex<-cex%>%mutate(fam_income_level=as_factor(as.character(fam_income_level)))

cex_sum<-cex%>%group_by(fam_income_level)%>%summarize(groc_fam=mean(grocery_food))

cex_sum<-cex_sum%>%mutate(fam_income_level=as.factor(as.character(fam_income_level)))

gg<-ggplot(cex_sum,aes(x=fct_reorder(.f=as.factor(fam_income_level),groc_fam),y=groc_fam,fill=fam_income

gg<-gg+geom_bar(stat="identity")

gg<-gg+xlab("Family Income Level")+ylab("Spending on Groceries")+scale_fill_discrete(name="Family Income
##Print
gg
```



Another reason why families spend more on groceries is becuase they have more children, particularly more older children. Teenagers typically add a considerable amount to food grocery bills. The figure below shows that families with no kids spend the least on groceries, while families with older teenagers spend the most.

Spending on Groceries by Family Structure

```
cex_sum<-cex%>%group_by(childage)%>%summarize(groc_fam=mean(grocery_food))

gg<-ggplot(cex_sum,aes(x=fct_reorder(.f=as.factor(childage),groc_fam),y=groc_fam,fill=fct_reorder(.f=as

gg<-gg+geom_bar(stat="identity")
##Print
gg<-gg+xlab("Family Structure")+ylab("Spending on Groceries") +scale_fill_discrete("Family Structure")

gg<-gg+theme(axis.text.x=element_blank())+theme(legend.position="bottom")

gg
```



Predictive Model

```
mod<-lm(log(grocery_food+1)~
          inc_rank+
          childage,
        data=cex
)

stargazer(mod)
```

Our predictive model combines the insights above. First, we use a log transform of grocery spending to

reflect the exponential distribtuion of the variable. Second, we use both family income and family structure as predictors in a linear regression. Our results indicate that for every one percentile change in family income, the family will spend 1 percent more on groceries. In addition, families whose oldest children are between 12 and 17 and who have one child less than 12 spend 70 percent more on groceries than families who have no children.

Accuracy of the Model

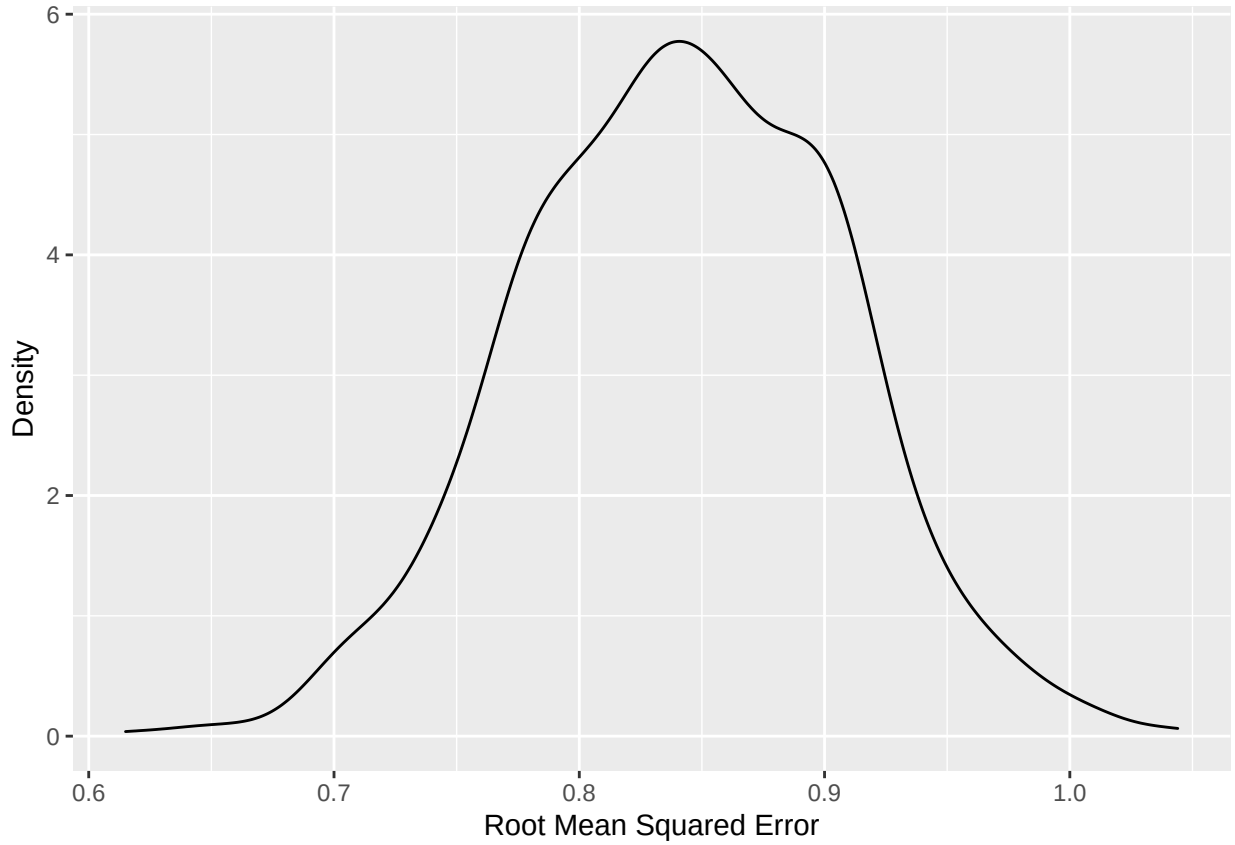
As it stands, we have only predicted 18 percent of the variance in grocery spending. Increase our budget and we will give you much better results.

Cross Validation

```
mod1_formula<-log(grocery_food+1)~
          inc_rank+
          childage


cex_cv<-cex%>%
  crossv_mc(n=1000,test=.2)


mod1_rmse_cv<-cex_cv %>%
  mutate(train = map(train, as_tibble)) %>% ## Convert to tibbles
  mutate(model = map(train, ~ lm(mod1_formula, data = .)))%>%
  mutate(rmse = map2_dbl(model, test, rmse))%>%
  select(.id, rmse) ## pull just id and rmse
```

Write in normal text.

```
#This is a comment
gg<-ggplot(mod1_rmse_cv,aes(x=rmse))
gg<-gg+geom_density()
gg<-gg+xlab("Root Mean Squared Error")+ylab("Density")
gg
```



The cross validation examines the extent to which the algorithm can predict out of sample data. It reruns the model 1,000 times, training it on on eighty percent of the data and testing it against twenty percent of the data each time. The minimum error from this analysis is about .64, the maximum about 1.05. The distribution of errors is symmetrical and centered on .85. The model appears to be consistently predictive of the outcome.

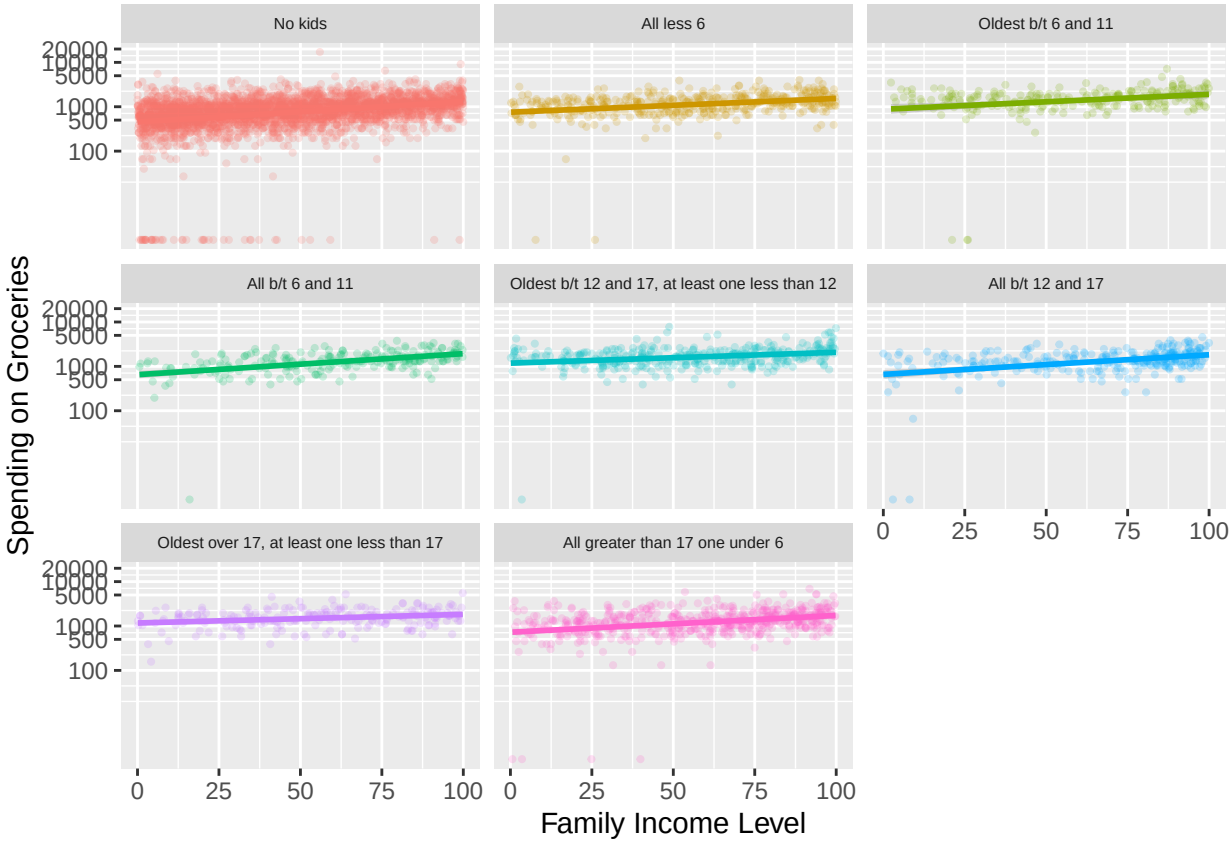
Predictions from this model

The figure below shows predicted grocery spending for families, by income rank and family structure.

Spending on Groceries by Family Income and Structure

```
gg<-ggplot(cex,aes(x=inc_rank,y=(grocery_food+1),color=childage))
gg<-gg+geom_point(size=.75,alpha=.2)
my.breaks=c(0,100,500,1000,5000,10000,20000)
#Change the scale
gg<-gg+scale_y_continuous(trans="log",breaks=my.breaks)
gg<-gg+geom_smooth(method="lm")
gg<-gg+facet_wrap(~childage)
gg<-gg+ theme(strip.text.x = element_text(size = 6))
gg<-gg+guides(color=FALSE)
gg<-gg+xlab("Family Income Level")+ylab("Spending on Groceries")
gg
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Recommendations

Based on this analysis, we have the following findings:

- Families who are above the 75th percentile in income generally will spend at least $1,000 on groceries.
- Families with children between 12 and 17 who are above the 25th percentile will spend at least $1,000 on groceries.
- Families with no children who are below the 50th percentile in income will very rarely spend $1,000 on groceries.

Our recommendation is to target advertising on the following family types:

- Families with incomes above the 75th percentile
- Families with children between 12 and 17 who are above the 25th percentile in income