CITY ANALYSIS

Brody Reid

March 13, 2020

# Analysis

# Combine data for city datasets.

### TORONTO

toronto\_full <- data.frame(cbind(toronto\_income$Income, toronto\_activity$Activity, toronto\_movements$Movements, toronto\_employed$Employed, toronto\_gdp$GDP))

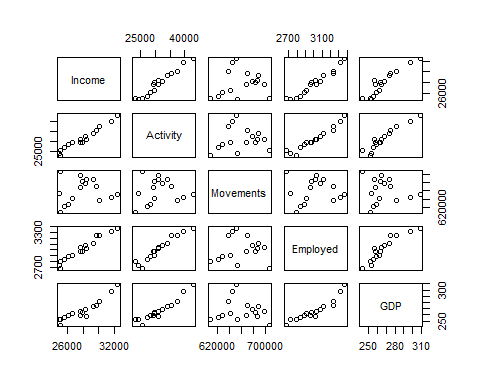
names(toronto\_full)[1] <- "Income"  
names(toronto\_full)[2] <- "Activity"  
names(toronto\_full)[3] <- "Movements"  
names(toronto\_full)[4] <- "Employed"  
names(toronto\_full)[5] <- "GDP"

Define RMSE function.

RMSE <- function(x){  
 sqrt(c(sum(x^2))/length(x))  
}

Matrix plot.

pairs(toronto\_full)



Income as response: 10-fold cross-validation RMSE for i-variable models, i = 1, …, 4.

set.seed(123)  
  
train\_control <- trainControl(method = "cv", number = 10)  
  
step\_models\_toronto <- train(Income ~ Activity + Movements + Employed + GDP, data = toronto\_full, method = "leapSeq", tuneGrid = data.frame(nvmax = 1:4), trControl = train\_control)

## Warning in nominalTrainWorkflow(x = x, y = y, wts = weights, info =  
## trainInfo, : There were missing values in resampled performance measures.

step\_models\_toronto$results

## nvmax RMSE Rsquared MAE RMSESD RsquaredSD MAESD  
## 1 1 627.2643 1 603.6263 167.5962 0 176.2020  
## 2 2 453.2039 1 434.7875 212.5408 0 206.7009  
## 3 3 436.6121 1 406.8868 185.0913 0 160.0868  
## 4 4 349.9008 1 316.4267 190.8654 0 166.9799

step\_models\_toronto$bestTune

## nvmax  
## 4 4

summary(step\_models\_toronto$finalModel)

## Subset selection object  
## 4 Variables (and intercept)  
## Forced in Forced out  
## Activity FALSE FALSE  
## Movements FALSE FALSE  
## Employed FALSE FALSE  
## GDP FALSE FALSE  
## 1 subsets of each size up to 4  
## Selection Algorithm: 'sequential replacement'  
## Activity Movements Employed GDP  
## 1 ( 1 ) " " " " "\*" " "  
## 2 ( 1 ) "\*" "\*" " " " "  
## 3 ( 1 ) " " "\*" "\*" "\*"  
## 4 ( 1 ) "\*" "\*" "\*" "\*"

Therefore, the best model is one with all 4 variables. Model with all 4 variables.

model\_toronto <- lm(Income ~ Activity + Movements + Employed + GDP, data = toronto\_full)  
  
summary(model\_toronto)

##   
## Call:  
## lm(formula = Income ~ Activity + Movements + Employed + GDP,   
## data = toronto\_full)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -677.7 -198.4 11.1 221.2 496.2   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -7.037e+03 7.074e+03 -0.995 0.34121   
## Activity 1.077e-01 9.608e-02 1.121 0.28618   
## Movements 8.890e-03 4.218e-03 2.108 0.05882 .   
## Employed 6.128e+00 1.788e+00 3.428 0.00564 \*\*  
## GDP 2.703e+01 2.270e+01 1.191 0.25871   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 366.4 on 11 degrees of freedom  
## Multiple R-squared: 0.9819, Adjusted R-squared: 0.9754   
## F-statistic: 149.5 on 4 and 11 DF, p-value: 1.651e-09

lm.beta(model\_toronto)

## Activity Movements Employed GDP   
## 0.2535449 0.1099920 0.5461230 0.1978190

RMSE(model\_toronto$residuals)

## [1] 303.8373

AIC(model\_toronto)

## [1] 240.3338

Employment as response.

model\_emp\_toronto <- lm(Employed ~ Activity + Movements + Income + GDP, data = toronto\_full)  
summary(model\_emp\_toronto)

##   
## Call:  
## lm(formula = Employed ~ Activity + Movements + Income + GDP,   
## data = toronto\_full)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -61.930 -24.798 -6.564 9.461 80.760   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.583e+03 7.228e+02 2.190 0.05095 .   
## Activity 8.612e-03 1.161e-02 0.742 0.47369   
## Movements -8.890e-04 5.213e-04 -1.705 0.11615   
## Income 8.429e-02 2.459e-02 3.428 0.00564 \*\*  
## GDP -2.259e+00 2.745e+00 -0.823 0.42807   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 42.98 on 11 degrees of freedom  
## Multiple R-squared: 0.9687, Adjusted R-squared: 0.9574   
## F-statistic: 85.19 on 4 and 11 DF, p-value: 3.347e-08

RMSE(model\_emp\_toronto$residuals)

## [1] 35.63508

AIC(model\_emp\_toronto)

## [1] 171.7526

Activity prediction.

toronto\_newdf1 <- data.frame(Activity = 25601, Movements = 706390, Income = 25000, GDP = 243.08)  
toronto\_newdf2 <- data.frame(Activity = 28161, Movements = 706390, Income = 25000, GDP = 243.08)

predict(model\_emp\_toronto, toronto\_newdf1)

## 1   
## 2733.83

predict(model\_emp\_toronto, toronto\_newdf2)

## 1   
## 2755.877

A 10% change in Activity yields a 0.81% increase in employment.

### VANCOUVER

vancouver\_full <- data.frame(cbind(vancouver\_income$Income, vancouver\_activity$Activity, vancouver\_movements$Movements, vancouver\_employed$Employed, vancouver\_gdp$GDP))

names(vancouver\_full)[1] <- "Income"  
names(vancouver\_full)[2] <- "Activity"  
names(vancouver\_full)[3] <- "Movements"  
names(vancouver\_full)[4] <- "Employed"  
names(vancouver\_full)[5] <- "GDP"

10-fold cross-validation to determine best model.

set.seed(123)  
  
train\_control <- trainControl(method = "cv", number = 10)  
  
step\_models\_vancouver <- train(Income ~ Activity + Movements + Employed + GDP, data = vancouver\_full, method = "leapSeq", tuneGrid = data.frame(nvmax = 1:4), trControl = train\_control)

## Warning in nominalTrainWorkflow(x = x, y = y, wts = weights, info =  
## trainInfo, : There were missing values in resampled performance measures.

step\_models\_vancouver$results

## nvmax RMSE Rsquared MAE RMSESD RsquaredSD MAESD  
## 1 1 778.5557 1 711.1617 348.1135 0 337.7492  
## 2 2 675.2007 1 606.6415 446.1566 0 395.1312  
## 3 3 930.6042 1 870.1453 520.2241 0 494.0306  
## 4 4 893.9887 1 806.0213 499.6011 0 434.4745

step\_models\_vancouver$bestTune

## nvmax  
## 2 2

summary(step\_models\_vancouver$finalModel)

## Subset selection object  
## 4 Variables (and intercept)  
## Forced in Forced out  
## Activity FALSE FALSE  
## Movements FALSE FALSE  
## Employed FALSE FALSE  
## GDP FALSE FALSE  
## 1 subsets of each size up to 2  
## Selection Algorithm: 'sequential replacement'  
## Activity Movements Employed GDP  
## 1 ( 1 ) " " " " " " "\*"  
## 2 ( 1 ) " " "\*" " " "\*"

Full model with income as response.

model\_vancouver <- lm(Income ~ Activity + Movements + Employed + GDP , data = vancouver\_full)  
summary(model\_vancouver)

##   
## Call:  
## lm(formula = Income ~ Activity + Movements + Employed + GDP,   
## data = vancouver\_full)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1328.9 -294.6 142.2 367.2 1134.2   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 2801.55385 9143.41275 0.306 0.765  
## Activity 0.40434 0.68562 0.590 0.567  
## Movements -0.02384 0.02000 -1.192 0.258  
## Employed 9.50161 15.65915 0.607 0.556  
## GDP 258.46401 181.21387 1.426 0.182  
##   
## Residual standard error: 830.6 on 11 degrees of freedom  
## Multiple R-squared: 0.9587, Adjusted R-squared: 0.9437   
## F-statistic: 63.81 on 4 and 11 DF, p-value: 1.535e-07

RMSE(model\_vancouver$residuals)

## [1] 688.6774

AIC(model\_vancouver)

## [1] 266.5188

Model with employed as response.

model\_emp\_vancouver <- lm(Employed ~ Activity + Movements + Income + GDP, data = vancouver\_full)  
summary(model\_emp\_vancouver)

##   
## Call:  
## lm(formula = Employed ~ Activity + Movements + Income + GDP,   
## data = vancouver\_full)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -24.7412 -6.4571 -0.4326 7.0039 28.2792   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.312e+02 1.694e+02 -0.775 0.4548   
## Activity -2.423e-02 1.098e-02 -2.207 0.0495 \*  
## Movements 6.353e-04 3.541e-04 1.794 0.1003   
## Income 3.409e-03 5.618e-03 0.607 0.5563   
## GDP 6.265e+00 3.223e+00 1.944 0.0779 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 15.73 on 11 degrees of freedom  
## Multiple R-squared: 0.6514, Adjusted R-squared: 0.5246   
## F-statistic: 5.138 on 4 and 11 DF, p-value: 0.01394

RMSE(model\_emp\_vancouver$residuals)

## [1] 13.04374

AIC(model\_emp\_vancouver)

## [1] 139.5919

Activity prediction.

vancouver\_newdf1 <- data.frame(Activity = 13817, Movements = 379177, Income = 22300, GDP = 73.94)  
vancouver\_newdf2 <- data.frame(Activity = 15199, Movements = 379177, Income = 22300, GDP = 73.94)

predict(model\_emp\_vancouver, vancouver\_newdf1)

## 1   
## 314.1195

predict(model\_emp\_vancouver, vancouver\_newdf2)

## 1   
## 280.6313

Therefore a 10% increase in activity produces an 11% decrease in employment.

### MONTREAL

montreal\_full <- data.frame(cbind(montreal\_income$Income, montreal\_activity$Activity, montreal\_movements$Movements, montreal\_employed$Employed, montreal\_gdp$GDP))

names(montreal\_full)[1] <- "Income"  
names(montreal\_full)[2] <- "Activity"  
names(montreal\_full)[3] <- "Movements"  
names(montreal\_full)[4] <- "Employed"  
names(montreal\_full)[5] <- "GDP"

10-fold cross-validation to determine best model.

set.seed(123)  
  
train\_control <- trainControl(method = "cv", number = 10)  
  
step\_models\_montreal <- train(Income ~ Activity + Movements + Employed + GDP, data = montreal\_full, method = "leapSeq", tuneGrid = data.frame(nvmax = 1:4), trControl = train\_control)

## Warning in nominalTrainWorkflow(x = x, y = y, wts = weights, info =  
## trainInfo, : There were missing values in resampled performance measures.

step\_models\_montreal$results

## nvmax RMSE Rsquared MAE RMSESD RsquaredSD MAESD  
## 1 1 819.9392 1 804.7543 434.4762 0 419.7102  
## 2 2 463.1171 1 430.6188 269.3607 0 267.2049  
## 3 3 812.5270 1 772.7510 208.0452 0 224.0422  
## 4 4 581.4217 1 529.0295 346.8912 0 325.7690

step\_models\_montreal$bestTune

## nvmax  
## 2 2

summary(step\_models\_montreal$finalModel)

## Subset selection object  
## 4 Variables (and intercept)  
## Forced in Forced out  
## Activity FALSE FALSE  
## Movements FALSE FALSE  
## Employed FALSE FALSE  
## GDP FALSE FALSE  
## 1 subsets of each size up to 2  
## Selection Algorithm: 'sequential replacement'  
## Activity Movements Employed GDP  
## 1 ( 1 ) "\*" " " " " " "  
## 2 ( 1 ) "\*" " " " " "\*"

Full model with income as response.

model\_montreal <- lm(Income ~ Activity + Movements + Employed + GDP , data = montreal\_full)  
summary(model\_montreal)

##   
## Call:  
## lm(formula = Income ~ Activity + Movements + Employed + GDP,   
## data = montreal\_full)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -723.50 -225.08 -77.17 400.30 560.18   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.486e+04 7.774e+03 -1.912 0.082275 .   
## Activity 8.655e-01 1.655e-01 5.229 0.000281 \*\*\*  
## Movements 8.005e-03 7.236e-03 1.106 0.292175   
## Employed 1.387e+00 1.001e+01 0.139 0.892305   
## GDP 1.959e+02 4.758e+01 4.116 0.001711 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 459.4 on 11 degrees of freedom  
## Multiple R-squared: 0.9875, Adjusted R-squared: 0.9829   
## F-statistic: 216.4 on 4 and 11 DF, p-value: 2.242e-10

RMSE(model\_montreal$residuals)

## [1] 380.8803

AIC(model\_montreal)

## [1] 247.5656

Model with employed as response.

model\_emp\_montreal <- lm(Employed ~ Activity + Movements + Income + GDP, data = montreal\_full)  
summary(model\_emp\_montreal)

##   
## Call:  
## lm(formula = Employed ~ Activity + Movements + Income + GDP,   
## data = montreal\_full)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -20.448 -12.123 5.343 8.524 17.200   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.279e+02 2.181e+02 2.421 0.034 \*  
## Activity 4.158e-03 9.213e-03 0.451 0.661   
## Movements 4.383e-05 2.291e-04 0.191 0.852   
## Income 1.256e-03 9.064e-03 0.139 0.892   
## GDP 2.129e+00 2.190e+00 0.972 0.352   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 13.82 on 11 degrees of freedom  
## Multiple R-squared: 0.8733, Adjusted R-squared: 0.8272   
## F-statistic: 18.96 on 4 and 11 DF, p-value: 6.743e-05

RMSE(model\_emp\_montreal$residuals)

## [1] 11.4602

AIC(model\_emp\_montreal)

## [1] 135.4502

Montreal activity prediction.

montreal\_newdf1 <- data.frame(Activity = 9447, Movements = 423155, Income = 22000, GDP = 126.36)  
montreal\_newdf2 <- data.frame(Activity = 10392, Movements = 423155, Income = 22000, GDP = 126.36)

predict(model\_emp\_montreal, montreal\_newdf1)

## 1   
## 882.3694

predict(model\_emp\_montreal, montreal\_newdf2)

## 1   
## 886.2987

Therefore a 10% increase in activity produces a 0.45% increase in employment.