The Effects of Gross Square Footage and Estimated Expense on the Full Market Value of New York City Condominiums.

Jaan Choudhri

Abstract

As housing begins to fill up and more people obtain increasing amounts of income, investors will begin to invest in condomiums in densely located cities such as New York City and Los Angeles. In order to determine which characteristics should be taken under consideration when investing in condominiums in New York City, we created a linear regression model to assess the impact of various predictors on the full market values of condos. We discovered that estimated gross income and gross squared footage can affect the full market value of New York City condos.

Introduction

As a means of being ahead of inflation, people will invest in various things, such as stocks, businesses, and housing. Large cities, particularly New York City, has a plethora of condominiums that would a worthwile investment, particularly due to need for housing all across the boroughs of New York for an ever-increasing population. Unfortunately, investors struggle to determine what should be prioritized when deciding on an investment. Evidence suggests that the market value of a condo can be determined by the estimated gross income, borough ID, or neighborhood. This struggle has continued due to a lack of quantifying the influence of various characteristics of the condos. In order to discover which factors hold influence over the market value of a condo in New York, we've developed various regression models and have decided to select an optimal model that can be useful in deciding which NYC condos are worth an investment. In our model, we were interested in examining the effect of estimated gross income, which is the estimated income per SquareFoot multiplied by gross squared footage, and gross squared footage, which is the total area inside of a building, on New York City's condominium full market value.

Method of model design

We first implement a bivariate test for a simple linear regression model in order to determine the relationship between the various factors and full market value of the condominiums individually.

```r
library(tidyverse)
```

```
## -- Attaching packages ------------------------------------------------------------------------------
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.0     v dplyr   1.0.4
## v tidyr   1.1.1     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.1
```

```
## -- Conflicts ----------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
nyc <- read.csv("nyc-condos.csv")
model <- lm(FullMarketValue ~ EstGrossIncome, data = nyc)
summary(model)
```

```
##
## Call:
## lm(formula = FullMarketValue ~ EstGrossIncome, data = nyc)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -27512733   -401244    860703   1400341  24988987
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -1.647e+06  3.688e+05  -4.466 1.34e-05 ***
## EstGrossIncome  5.990e+00  6.069e-02  98.710  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4518000 on 198 degrees of freedom
## Multiple R-squared:  0.9801, Adjusted R-squared:   0.98
## F-statistic:  9744 on 1 and 198 DF,  p-value: < 2.2e-16
```

From the data table above, the p-value is 2.0e-16, which is very small. Therefore, we can say that the estimated gross income is strongly related to the full market value of condos without any conditional factors.

```r
model <- lm(FullMarketValue ~ GrossSqFt, data = nyc)
summary(model)
```

```
##
## Call:
## lm(formula = FullMarketValue ~ GrossSqFt, data = nyc)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -95596170  -1250348   1636134   3543727  72561306
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.209e+06  1.231e+06  -3.419 0.000762 ***
## GrossSqFt    2.514e+02  8.757e+00  28.709  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14090000 on 198 degrees of freedom
## Multiple R-squared:  0.8063, Adjusted R-squared:  0.8053
## F-statistic: 824.2 on 1 and 198 DF,  p-value: < 2.2e-16
```

From the data table above, the p-value is 2.2e-16, which is very small. Therefore, we can say that the gross square footage of the condos are strongly related to the full market value of condos without any conditional factors.

```r
model <- lm(FullMarketValue ~ GrossSqFt + EstGrossIncome + EstGrossIncome*GrossSqFt, data = nyc)
summary(model)
```

```
##
## Call:
## lm(formula = FullMarketValue ~ GrossSqFt + EstGrossIncome + EstGrossIncome *
##     GrossSqFt, data = nyc)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -17172416   -657357   -183800    634613  13521513
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               9.993e+05  3.157e+05   3.165   0.0018 **
## GrossSqFt                -6.515e+01  5.546e+00 -11.748  < 2e-16 ***
## EstGrossIncome            6.433e+00  1.609e-01  39.990  < 2e-16 ***
## GrossSqFt:EstGrossIncome  1.709e-06  2.009e-07   8.507 4.56e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3123000 on 196 degrees of freedom
## Multiple R-squared:  0.9906, Adjusted R-squared:  0.9904
## F-statistic:  6873 on 3 and 196 DF,  p-value: < 2.2e-16
```

The model above is our final model used to predict the full market value of the condos. Our model contains three predictors: the estimated gross income, the gross square footage of the condos, and the interaction term between these two characteristics. As seen in the table above, the p-value for our three predictors is 2.0e-16, 2.0e-16, and 4.56e-15 respectively, indicating that all of our predictors are strongly influential. Our adjusted R-Squared is .9906, meaning that 99.06% of the full market values can be explained by our model. Our model is a strong predictor for full market value of the condos.

```r
resample <- function(data, resample_size) {
  row_numbers <- sample.int(nrow(data), size=resample_size, replace = TRUE)
  resample_data <- data[row_numbers, ]
  return(resample_data)
}

set.seed(42069)

Results <- replicate(1000, lm(FullMarketValue ~ GrossSqFt + EstGrossIncome + EstGrossIncome*GrossSqFt, 
Boot_Res <- data.frame(t(Results))
summary(Boot_Res)
```

```
##   X.Intercept.       GrossSqFt       EstGrossIncome   GrossSqFt.EstGrossIncome
##  Min.   :-465927   Min.   :-118.92   Min.   :5.105   Min.   :-3.842e-06
##  1st Qu.: 671685   1st Qu.: -72.11   1st Qu.:6.150   1st Qu.: 1.231e-06
##  Median : 913291   Median : -63.72   Median :6.447   Median : 1.676e-06
##  Mean   : 881205   Mean   : -62.83   Mean   :6.453   Mean   : 1.533e-06
##  3rd Qu.:1108311   3rd Qu.: -53.73   3rd Qu.:6.751   3rd Qu.: 2.024e-06
##  Max.   :1979290   Max.   : -12.78   Max.   :7.667   Max.   : 3.119e-06
```

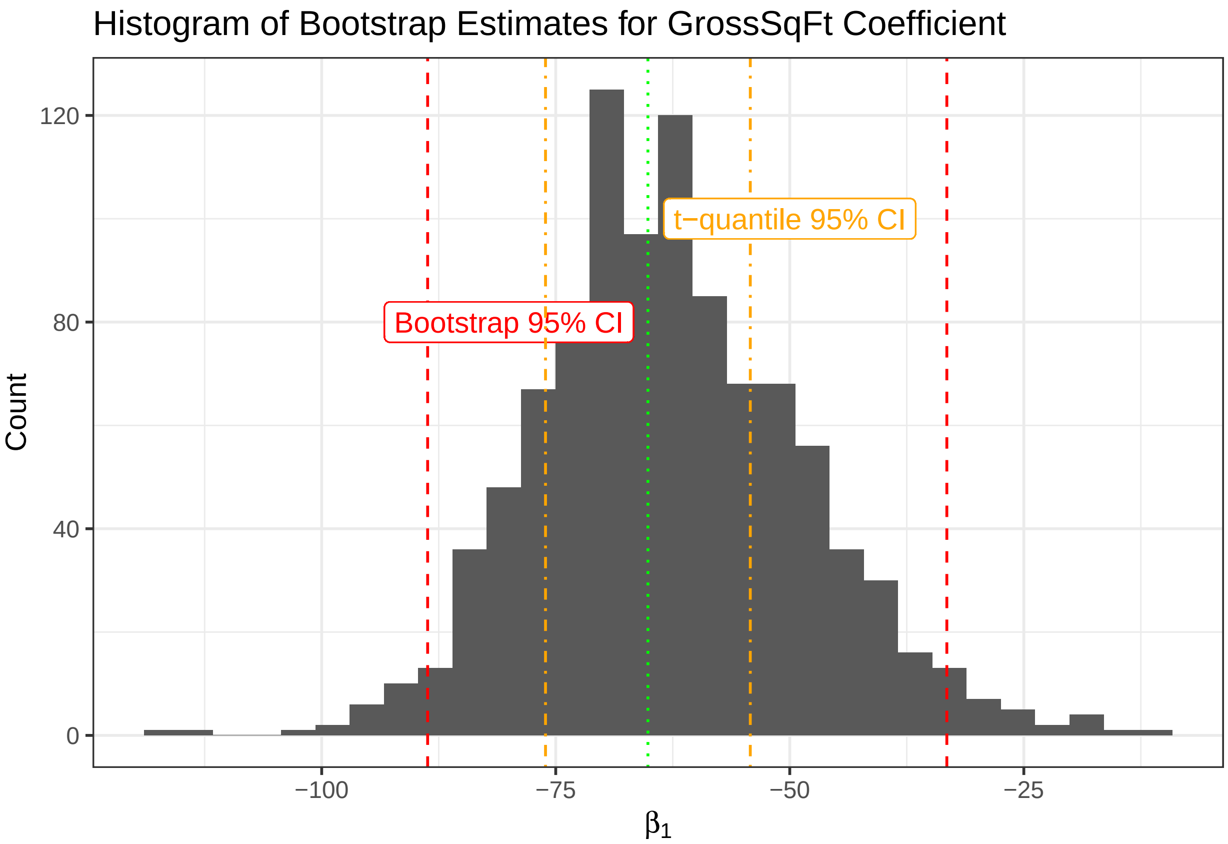
```r
Area_CI <- quantile(Boot_Res$GrossSqFt, probs = c(0.025, 0.975))
Income_CI <- quantile(Boot_Res$EstGrossIncome, probs = c(0.025, 0.975))
Area_CI
```

```
##      2.5%     97.5%
## -88.68105 -33.22267
```

```
BS_Ests <- Boot_Res %>% ggplot(aes(x = GrossSqFt)) +
  geom_histogram() +
  geom_vline(xintercept = model$coefficients["GrossSqFt"],
             col = "green", lty = "dotted") +
  labs(title = "Histogram of Bootstrap Estimates for GrossSqFt Coefficient",
       x = expression(beta[1]),
       y = "Count") +
  theme_bw()

BS_Ests <- BS_Ests +
    geom_vline(xintercept = quantile(Boot_Res$GrossSqFt, probs=c(0.025,0.975)),
    col = "red", lty = "dashed") +
  geom_label(x = -80, y = 80, label = "Bootstrap 95% CI", col = "red")
BS_Ests <- BS_Ests +
    geom_vline(xintercept = confint(model)["GrossSqFt",], col = "orange", lty = "dotdash") +
  geom_label(x = -50, y = 100, label = "t-quantile 95% CI", col = "orange")
BS_Ests
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



In this bootstrap model, we discovered that the coefficient of the gross square footage has serious deviation from our model. In the bootstrap assumption, we are 95% confident that the coeffcent for gross square footage is -88.68105 and -33.22267. The mean value derived from our bootstrap model, 22.55, is vastly different from the original coefficient value of -65.153. In this case, we will use the original coefficient in our model.

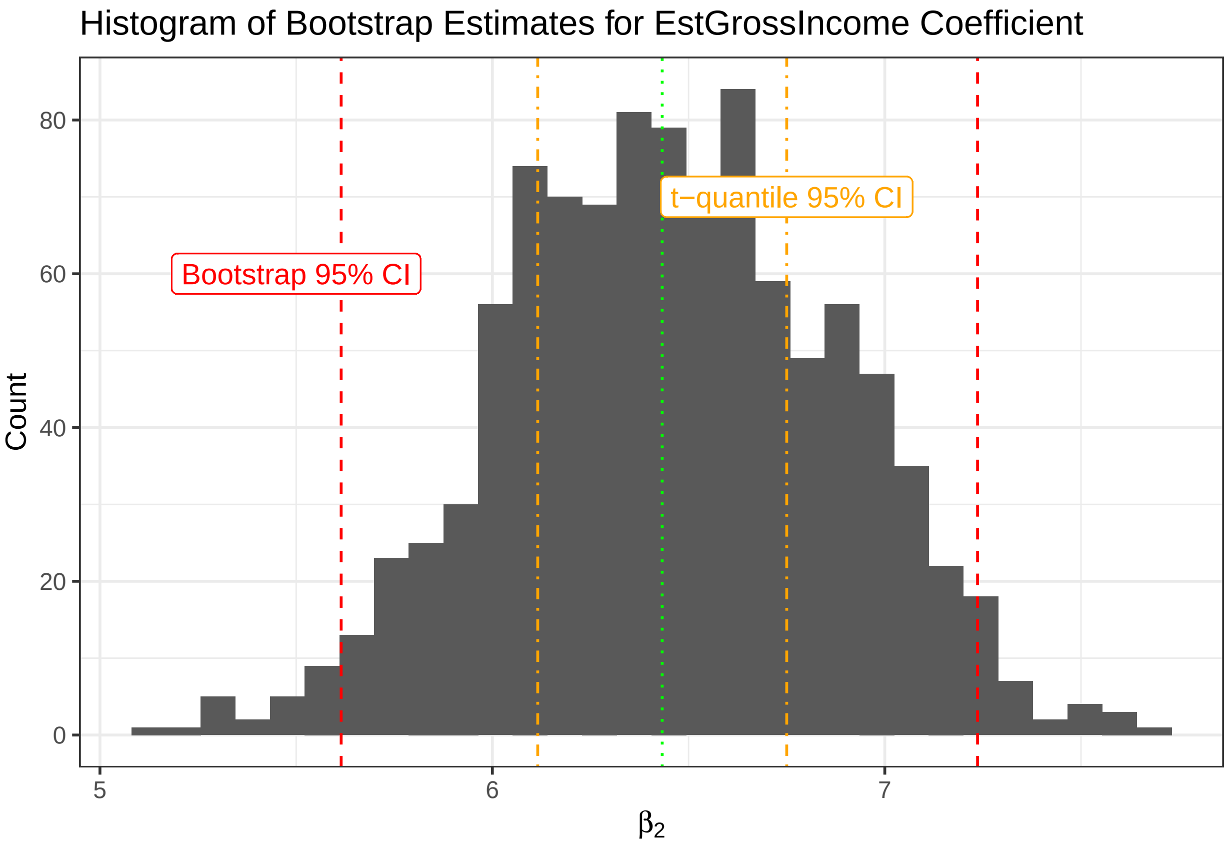
```
Income_CI
```

```
##      2.5%     97.5%
## 5.614694 7.236401
```

```
Area_CI <- quantile(Boot_Res$GrossSqFt, probs = c(0.025, 0.975))
Income_CI <- quantile(Boot_Res$EstGrossIncome, probs = c(0.025, 0.975))

BS_Ests <- Boot_Res %>% ggplot(aes(x = EstGrossIncome)) +
  geom_histogram() +
  geom_vline(xintercept = model$coefficients["EstGrossIncome"],
             col = "green", lty = "dotted") +
  labs(title = "Histogram of Bootstrap Estimates for EstGrossIncome Coefficient",
       x = expression(beta[2]),
       y = "Count") +
  theme_bw()
BS_Ests <- BS_Ests +
    geom_vline(xintercept = quantile(Boot_Res$EstGrossIncome, probs=c(0.025,0.975)),
    col = "red", lty = "dashed") +
  geom_label(x = 5.5, y = 60, label = "Bootstrap 95% CI", col = "red")
BS_Ests <- BS_Ests +
    geom_vline(xintercept = confint(model)["EstGrossIncome",], col = "orange", lty = "dotdash") +
  geom_label(x = 6.75, y = 70, label = "t-quantile 95% CI", col = "orange")
BS_Ests
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



After implementing our bootstrap method, we found that the coefficient for the estimated gross income deviates from our original model. In our bootstrap model, we are 95% confident that the coefficient for the estimated gross income would be between 5.615 and 7.236. However, the mean coefficient from the bootstrap assumption, 3.3755, is not in our new confidence interval and vastly different from the original case, which has a mean coefficient of 6.433. In this instance, we can use our original coefficient in our model.

```
Income_Res <-with(Boot_Res, c(mean=mean(EstGrossIncome),
                              median = median(EstGrossIncome),
                              variance = var(EstGrossIncome),
                              stdev = sd(EstGrossIncome)))
Area_Res <-with(Boot_Res, c(mean=mean(GrossSqFt),
                              median = median(GrossSqFt),
                              variance = var(GrossSqFt),
                              stdev = sd(GrossSqFt)))
summary(Area_Res)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -63.72  -63.05  -24.30   22.55   61.30  202.52
```

Discussion

Our results show that when adjusted for gross squared footage, the relationship of full market value of NYC condos and estimated gross income is positive.

With the extremely high coefficient of 0.9906 in our model, there is a possibility of using our model to predict the full market value of condos in NYC using two predictors, the gross squared footage and estimated gross income. This model may be limited by a outlier value that has a very small area. Removing this outlier may help make our model more accurate.

Conclusion

In our study, we have examined the relationship between the full market value of New York City condos and various characteristics of the condos, including gross square footage and estimated gross income. Our linear regression model shows that gross square footage combined with estimated gross income can be influential on the market value. For example, when examining higher estimated gross income, the full market value of the condo is more expensive. Therefore, estimated gross income can determine the market value of NYC condos. With these results, investors can discover which characteristics of condominiums are important to making a profitable investment.