Simple Predictors of Homelessness in Colorado

Elizabeth Rodriguez
University of Colorado Boulder
May 1, 2022

Abstract
The purpose of this research and analysis was to determine the significance of predictors of homelessness in Colorado and to try to find a model that would best fit the data. The data spans ten years, from 2007 to 2016. I use the estimates of homeless individuals per year from The Department of Housing and Urban Development. Some of the methods for finding a model include Backwards Elimination, PCA, adjusted R-squared, GLS, and transformations.

Introduction
If you have lived in Colorado for as little as four years you have probably noticed the same issue as I have, an issue that seems to be worse and worse with every passing year. Homelessness is a pandemic in this country, growing with every passing year, and seemingly unstoppable. If you were to ask anyone in Colorado, maybe a passerby on the street, if they thought homelessness was worse than the year before, most people would probably say yes, I would. Yet, when looking at the overall number of homelessness in Colorado, it seems to have decreased in past years. What would cause this? Maybe a more important and broad question to ask is what could predict homelessness (in Colorado)? These are some of the questions I look to answer in this paper.

Methods and Results
Data and Cleaning
The dataset used was 'Homelessness' from Kaggle. This dataset originally had 6 columns of data including Year, States, CoC Number, CoC Name, Measures, and Count.

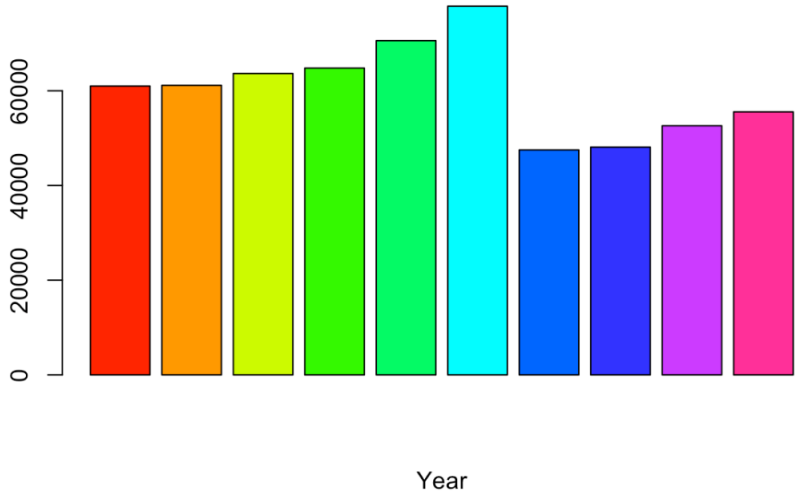
Year: 2007 - 2016
States: 50
CoC Number: Continuum of Care Number
CoC Name: Continuum of Care Name
Measures: Type of shelter/count of homeless individuals ex: homeless veterans, homeless families, etc.
Count: Number of homeless individuals for each measure in each state (per year)

The CoC Number and CoC Name was not used in the analysis. The data was over ten years, from 2007 to 2016. The data was loaded into R as homedata. Any rows with a value of 'NA'

were omitted immediately after loading in the data. The resulting dataset consisted of 86,529 rows, and three columns; Year, Measures, and Count.
To be able to find significant predictors of homelessness in Colorado, I summed every Count for each Measure, to find the total number of homeless individuals in Colorado per year. This vector was named ‘totalcnt’.
The population of Colorado was found by The United States Census Bureau (from 2007 to 2016). A vector was created from this information called ‘COpopulation’. The average salary in Colorado from 2007 to 2016 and made a vector named ‘COAvSal’. ‘UnempRate’ is a vector consisting of the unemployment rate in Colorado from 2007 to 2016, and ‘AvHousing’ is a vector consisting of the average housing prices in Colorado from 2007 to 2016. I then created a new data frame using COpopulation, COAvSal, UnempRate, AvHousing, Year, and totalcnt.

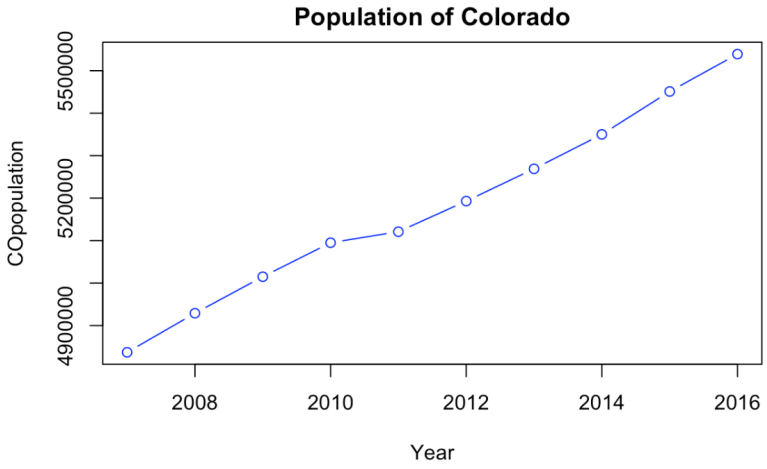
Analysis of Vectors

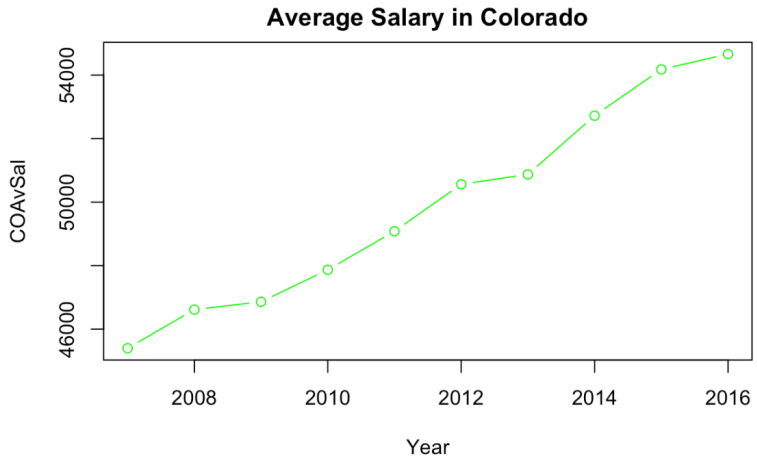
The vector ‘totalcnt’ was plotted into a bar graph to compare the total amount of homeless individuals in Colorado for each year in the analysis.



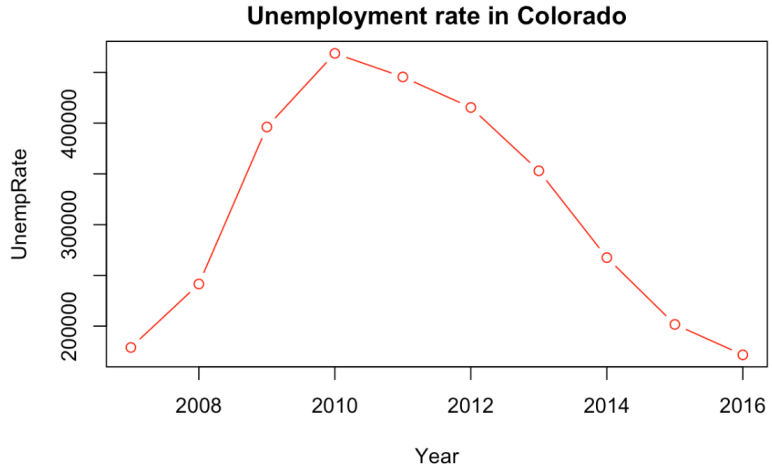
From 2007 to 2012 we see homelessness is increasing with each year. In 2012 homelessness in Colorado hit the maximum in this time period, of 77,868 individuals. We then see a considerable decrease in 2013, when the amount of homeless individuals is 47,506. What caused this decrease of thousands? The count of homeless individuals is based on three point-in-time (PIT) estimates including the Denver metro area, the Colorado Springs metro, and rural communities. In 2013 the BOS CoC (Balance of State Continuum of Care) changed the methodology of counting homeless individuals in Colorado by removing the count of homeless individuals in rural communities. Prior to 2013, the BOS CoC would count homeless individuals in Pueblo and use this number to estimate homelessness in other rural areas in Colorado. This change of methodology is the largest factor for the drop between 2012 and 2013.

The population vector ‘COpopulation’ is continually increasing each year, along with the average salary in Colorado vector ‘COAvSal’.

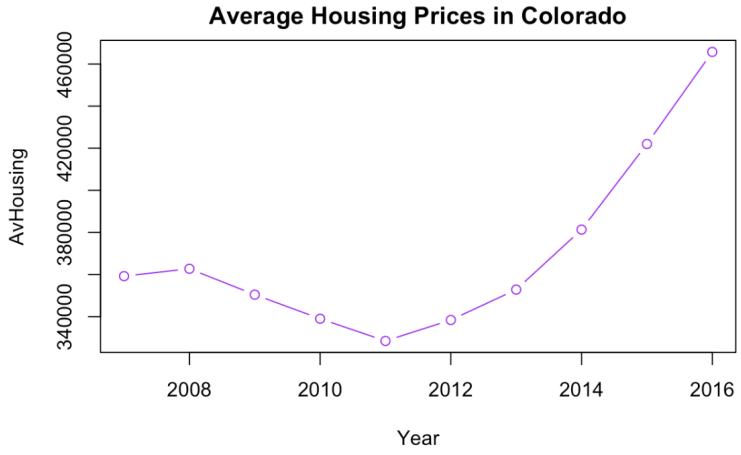




The graph that plots the unemployment rate in Colorado is not as simple, it seems to be increasing until 2010, then decreases from there.

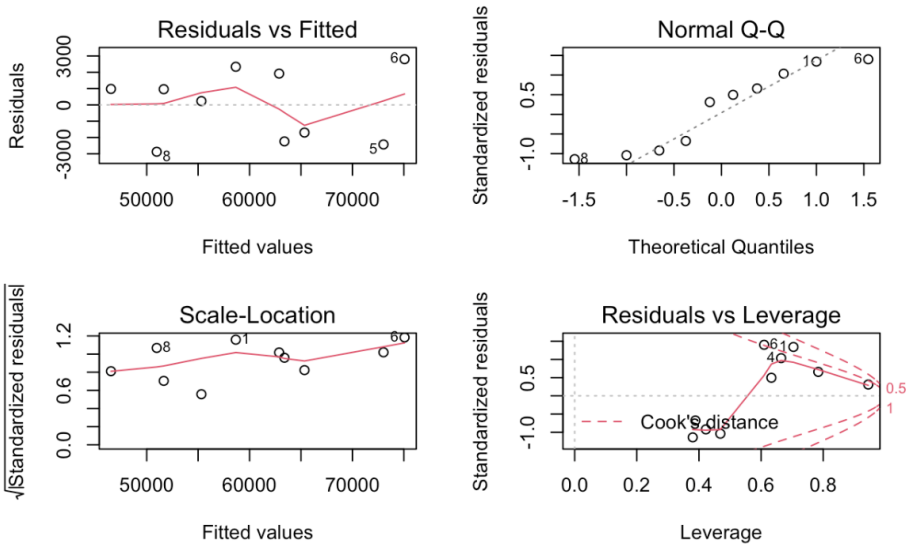


There is a large increase in unemployment from 2008 to 2009. This was caused by the recession that started between 2007 and 2008. Many people lost their jobs not just in Colorado but in The United States. After 2010 we see that it is continually decreasing, a result of the economy recovering from said recession.
The average housing prices in Colorado also were affected for the same reason. From 2007 to 2008 it is increasing, then after 2008, housing prices decrease until after 2011 when they start to increase once again. This is also due to the recession. The housing market was one of the industries that were part of the economic downfall and took a hit for multiple years.



Regression Analysis

I wanted to create the best multiple linear regression model. Usually, the 'best' model would mean the model with the lowest mean squared error, but in the case of this analysis, the best model is a combination of significant predictors, high adjusted R-squared value, and smaller error terms (mean squared error). I started with a multiple linear regression model using totalcnt as the independent variable and Year, COpopulation, COAvSal, UnempRate, and AvHousing as the covariates. This model is called lmod1. Looking at the summary of lmod1, only one of the dependent variables was not significant. Every other variable has a p-value of less than 0.05. This means that other than Year, all the other variables are significant in terms of predicting total homeless individuals in Colorado. The adjusted R-squared is 0.8904, meaning that about 89 percent of the variation of totalcnt is explained by the covariates used in the model. When plotting the diagnostic plots for lmod1, we see that the Residual vs Fitted shows some structure, but this is expected since we only have 10 data points (per each year looked at). The Normal Q-Q also is a result of a smaller data pool. The points loosely fall along the diagonal with some heavy tails.

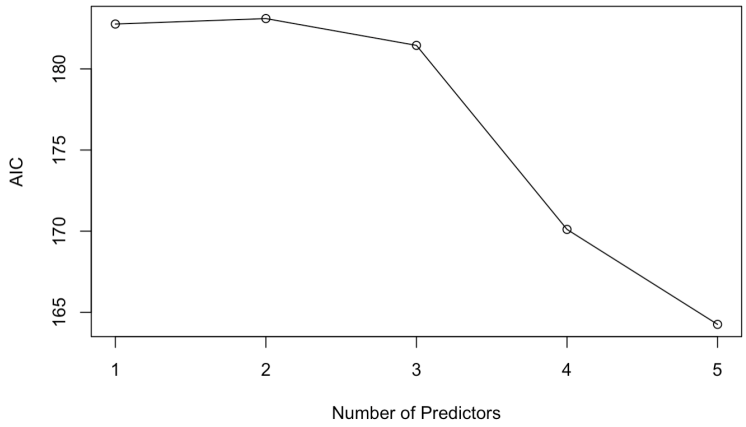


Although the adjusted R-squared is large and (most) predictors are significant, the model has high residuals, a residual standard error of 3200, and a mean squared error of 4096112 (yikes!). The goal is now to find a model with a smaller mean squared error and residual standard error that retains the significance of predictors and a large adjusted R-squared.

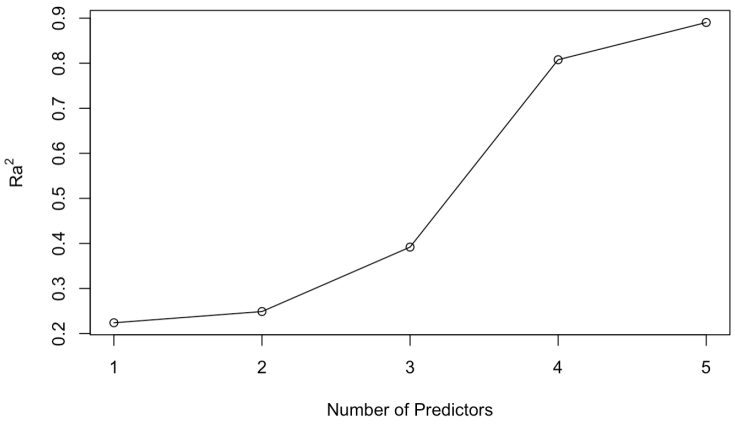
I first used Backwards Elimination: a process in which you remove predictors based on a significance level picked prior. I let alpha be 0.05 and removed any predictors that were not significant at this level. I fit the model with the same independent variable and dependent variables minus Year (due to insignificance/above alpha). This model is called lmodB1.

Although all predictors are significant, the adjusted R-squared came out to be 0.8077, smaller than in the first model. The residual standard error and the mean squared error were 4239 and 8983985 respectively. I rejected that this model was better than the first.

The next model selection used was AIC (Akaike Information Criterion). AIC is a criterion-based method that allows you to see how many predictors would produce the best model. Graphing this alone shows that using all five predictors in our model results in a large increase in the (log) likelihood at its maximum, meaning that using all five predictors is better than using any other number of predictors.



We could also check this by using another criterion-based method: Adjusted R-Squared. Adjusted R-squared always increases when a variable is added, whether that variable is significant or not. This is a weighted adjusted R-squared, so a predictor will only increase the adjusted R-squared value if it is significant or has some predictive value. We can see from the graph that the number of predictors with the highest adjusted R-squared value is all five.



From Backwards Elimination, AIC, and Adjusted R-squared I have found that the first model using all the variables was coming out to be the best model.
I then decided to use transformations to the covariates. I first used a square root transformation on the variable ‘Year’, since it was the only variable that was insignificant in the first model. A square root transformation can reduce heteroscedasticity of residuals (residual errors having different variances). This model was called lmodT1 and returned a residual standard error of 4002 and a mean squared error of 8006735. Although every prediction, now including the transformed Year variable, was significant, the errors were still quite high. I then tried a log transformation on the variable Year. The log transformation normalizes skewed data, (similar to the square root transformation). This model (lmodT2) returned almost exactly the same statistics

as lmod1, with an adjusted R-squared of .89, the same significant predictors, and a mean squared error of 4110844.
I also used GLS to transform the first model (lmod1). This model returned very similar results in regards to high errors and significance of predictors.
If there is a high correlation between two covariates then an interaction term could be able to take some of the colinearity out of the model and return a model with smaller residuals.

	Year1	COpopulation	totalcnt	COAvSal	UnempRate	AvHousing
Year1	1.0000000	0.9958134	-0.4385427	0.9924963	-0.2184211	0.6762503
COpopulation	0.9958134	1.0000000	-0.4465946	0.9886643	-0.2345662	0.7159105
totalcnt	-0.4385427	-0.4465946	1.0000000	-0.4389669	0.5568802	-0.5293764
COAvSal	0.9924963	0.9886643	-0.4389669	1.0000000	-0.3005461	0.7178555
UnempRate	-0.2184211	-0.2345662	0.5568802	-0.3005461	1.0000000	-0.7632493
AvHousing	0.6762503	0.7159105	-0.5293764	0.7178555	-0.7632493	1.0000000

There seemed to be a high correlation between many of the covariates. Specifically, Year and COpopulation, Year and COAvSal, and COpopulation and COAvSal all had a correlation of above .9.
The first model (lmod_interact1) used the interaction term between Year and COpopulation. This returned an adjusted R-squared value of .9197 and a residual standard error of 2739. The significance of almost every predictor changed, and now only COAvSal and UnempRate were the only two significant predictors. I wanted to make sure that most predictors would be significant so I tried another interaction term. The next model (lmod_interact2) used Year and CoAvSal as the interaction term. This model returned an adjusted R-squared value of 0.8923 and a residual standard error of 3172. The only significant predictors came out to be COpopulation, UnempRate, and Avhousing. Because the standard error was almost the same as the first model used in my analysis and the p-value wasn't much better I rejected this model and tried another interaction term. The next model (obviously called lmod_interact3) used the interaction term COpopulation and COAvSal. This model was almost the exact same as lmod_interact2 with a slightly smaller residual standard error (3092), but the significant predictors dwindled to only one, UnempRate.
Due to these results, I tried a combination of interaction terms and predictors, starting with a new model (lmod_interact4) I added two interaction terms; Year*COpopulation and Year*CoAvSal. This model produced the lowest mean squared yet, 2038527. The issue with this model was that the only predictor that came out significant was UnempRate. Next, I tried the same model but removed the Year variable (but not the interaction terms). This model named lmod_interact5, gave the best statistics thus far, with an adjusted R-squared of 0.9256, a mean squared error of 2781937, and almost all significant covariates, with the exception of Year*COpopulation. This seems to be the best model of fit.

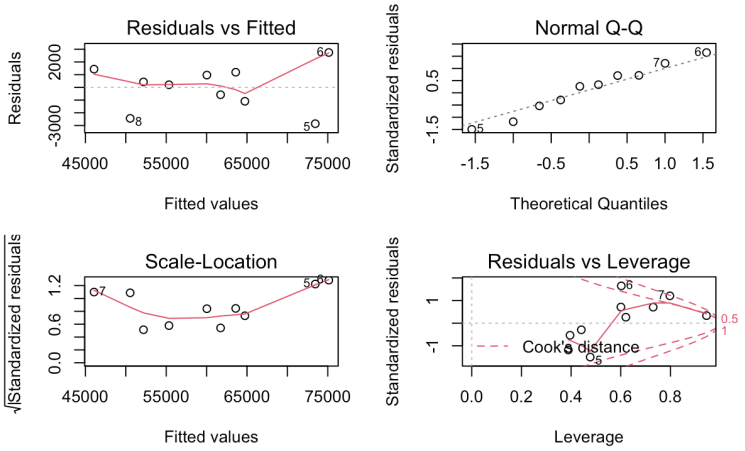
```
Call:
lm(formula = totalcnt ~ . + Year * COpopulation + COpopulation *
    COAvSal - Year - COAvSal, data = COpopdata)

Residuals:
      1       2       3       4       5       6       7       8       9      10
  967.6  -578.2 -1095.8  1191.4 -2857.7  2744.2  1433.4 -2439.5   428.7   205.9

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)           2.835e+06  4.685e+05   6.051  0.00376 **
COpopulation         -7.600e+00  2.734e+00  -2.780  0.04982 *
UnempRate             2.943e-01  3.421e-02   8.603  0.00100 **
AvHousing             7.828e-01  1.156e-01   6.770  0.00248 **
Year:COpopulation     3.380e-03  1.320e-03   2.560  0.06263 .
COpopulation:COAvSal  3.832e-06  6.147e-07   6.234  0.00337 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2637 on 4 degrees of freedom
Multiple R-squared:  0.9669,    Adjusted R-squared:  0.9256
F-statistic: 23.39 on 5 and 4 DF,  p-value: 0.004628

[1] 2781937
```



Discussion/Conclusions

The best model that I was able to find was lmod_interact5. This model not only had the best adjusted R-squared, but the lowest mean squared error and most of the predictors are significant. The most significant predictor is unemployment rate (UnempRate). This makes sense, the more people that are unemployed the more people cannot afford to live and therefore are at risk of homelessness. It turns out that in almost every model this was true. It also makes sense that in most models the Year variable was not significant (or the least significant), one, because it is just the index, and two, due to the dip from 2012 to 2013 in the total number of individuals that are homeless in Colorado (totalcnt) was not quite predictable.

Although this analysis went through a multitude of ways to create a good model for predicting homelessness in Colorado, there is still much to learn and analyze when it comes to an issue like this. Colorado is just one small piece of this problem, and even just looking at Colorado there are other factors that have implications for homelessness such as mental health, health care, gentrification, etc. Hopefully, Colorado is able to focus more on the problem in the future rather than hiding it by changing its methodology.

References

"All-Transactions House Price Index for Colorado." *FRED*, 22 Feb. 2022, https://fred.stlouisfed.org/series/COSTHPI.

"Average & Median Sale Price for a New Home." *Average and Median Cost for A New Home in The United States*, http://www.fedprimerate.com/new_home_sales_price_history.htm.

Love(X):, Def. "Homelessness." *Kaggle*, 7 Aug. 2017, https://www.kaggle.com/datasets/adamschroeder/homelessness.

Published by Statista Research Department, and Jan 20. "Colorado: Average Annual Pay 2019." *Statista*, 20 Jan. 2021, https://www.statista.com/statistics/594366/colorado-average-annual-pay/.

Published by Statista Research Department, and Mar 9. "Colorado - Unemployment Rate 2021." *Statista*, 9 Mar. 2022, https://www.statista.com/statistics/189397/unemployment-rate-in-colorado-since-1992/.

Recession BLS Spotlight - Bureau of Labor Statistics. https://www.bls.gov/spotlight/2012/recession/pdf/recession_bls_spotlight.pdf.

Wilcox, Author: Katie. "Colorado's Homeless Population Is Increasing, so Why Does This HUD Report Say Otherwise?" *KUSA.com*, 7 Dec. 2017, https://www.9news.com/article/news/colorados-homeless-population-is-increasing-so-why-does-this-hud-report-say-otherwise/73-497483910.