# Predictive Models of the Amount of Physical Activity Across the United States

Katelyn Tessier

University of Minnesota Duluth

University Honors Capstone Project

April 23rd, 2015

**Abstract**

How physical activity affects individuals' health across the United States is an important topic for the well-being of the country. The purpose of this study is to examine how certain factors influence the amount of physical activity performed by adults in each state. Using statistical analysis techniques, such as linear regression and model selection, a model of physical activity, including several covariates, is constructed. These covariates come from various categories, such as, geographic, financial and economic, residents' living habits, and residents' health. The final model, model (1), consists of four covariates which are *obesity*, *refund*, *uninsured* and *poverty*. This model can be used to predict how the percent of physically active adults changes as these four covariates change.

## Introduction

Factors that affect the amount of physical activity performed by individuals throughout the United States is a topic that needs to be discussed. Furthermore, how physical activity affects individuals' health across the United States is an important topic. According to Centers for Disease Control and Prevention (2014), less than half of Americans meet the *2008 Physical Activity Guidelines*, which states that adults, 18-64 years of age, need at least 2 hours and 30 minutes of moderate-intensity aerobic activity every week or 1 hour and 15 minutes of high-intensity aerobic activity every week, as well as 2 or more days of muscle strengthening activity. The benefits of physical activity are substantial. It controls weight, reduces the risk of cardiovascular disease, reduces the risk for type 2 diabetes and metabolic syndrome, increases the chance to live longer, improves the ability to do daily activities, strengthens bones and muscles, improves mental health and mood, and reduces the risk of some cancers (CDC, 2011).

The amount of physical activity one performs is affected by several factors. The purpose of this study is to examine how certain factors influence the amount of physical activity performed by adults in each state. Using statistical analysis techniques, such as linear regression and model selection, a predictive model of physical activity will be presented.

## Data Description

There are several variables that can affect the amount of physical activity, whether they are expected or not. The data in this study was collected from a variety of web sources listed in the

reference section. By collecting data on several variables for each state in the United States, a dataset for this study was compiled.

The dependent variable in this study is *exercise*. *Exercise* is the percent of physically active adults for each state in 2010. Physically active is defined as those who participate in "at least 150 minutes per week of moderate-intensity activity, or 75 minutes per week of vigorous-intensity activity, or a combination of moderate-intensity and vigorous-intensity activity (multiplied by two) totaling at least 150 minutes per week" (CDC, 2010). The reason for choosing *exercise* as the dependent variable is because the purpose of the study is to observe how it is affected by several other variables.

Possible covariates that are considered as potential contributors to the amount of physical activity include several categories. These categories include geographic covariates, financial and economic covariates, covariates that capture residents' living habits, and covariates related to residents' health. All of the variables selected for this study are pooled from different web sources, and documented from 2010 to 2013. Furthermore, each variable was collected for each state.

The geographic variables include *longitude*, *latitude*, *temp*, and *sun*. *Longitude* is the average longitude coordinate. *Latitude* is the average latitude coordinate. *Temp* is the average annual temperature in degrees Fahrenheit. *Sun* is the average number of hours of sun in a year.

The financial and economic variables include *UR*, *income*, *credit*, *refund*. *UR* is the unemployment rate for 2013. *Income* is the American Community Survey (ACS) median

household income in 2013. *Credit* is the average credit score from Credit Karma. *Refund* is the average tax refund for 2011.

The variables that capture residents' living habits include *poverty*, *snap*, and *internet*. *Poverty* is the average annual poverty rate from 2010 to 2012. *Snap* is the percent of the state population on supplemental nutrition assistance program (SNAP) in 2012. *Internet* is the percent of the state population with internet connection on multiple devices as of 2011.

The variables related to residents' health include *uninsured*, *restaurants*, and *obesity*. *Uninsured* is the percent of the state population without health insurance in 2013. *Restaurants* is the number of eating and drinking venues. *Obesity* is the prevalence of self-reported obesity for adults in 2013.

*Obesity* is expected to be a very important factor in the model, because it is related to the amount of physical activity one performs. Approximately 35% of adults in the United States are obese (CDC, 2014). This number changes to approximately 40% for adults 40-59 years old, 30% for adults 20-39 years old, but remains approximately the same for adults over 60 years old (CDC, 2014). The amount of adults in the United States who are obese is a surprisingly high percent. This is why it is important to analyze this factor in this study, and to see if it affects physical activity as expected.

**Methods**

Ordinary linear regression was used to analyze the data. *Longitude, latitude, income, temp, sun,*

*refund, credit,* and *restaurants* are standardized to have mean zero and unit variance. When interpreting standardized variables, it is based on one standard deviation difference in the variable, instead of one unit difference in the variable.

Once particular variables were scaled, ordinary linear regression was used again to analyze the data. The full model includes all of the variables mentioned in data description, as well as these re-scaled variables. To decide on more accurate models of physical activity than the full model, adjusted R-square, residual sum of squares, R-square, Mallow's Cp, Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) were examined. R-square is a measure of how close the data is to the regression line. The best model is the model with the smallest difference between the R-square of the current model and the R-square value of the model with an additional covariate. Adjusted R-square is similar to R-square, but it is a modified version that adjusts for the number of covariates in a model. The best model is one with the highest value. Residual sum of squares is a measure of the discrepancy between the data and the model. The best model is one with the lowest value. Mallow's Cp statistic compares the precision and bias of the full model with the selected models. The best model is the model with a small Cp and a Cp value that is less than the number of covariates. AIC is an estimate of a constant added to the relative distance between the unknown true likelihood function of the data and the likelihood function of the model ("AIC vs. BIC", 2007). The best model is the model with the lowest value. BIC is an estimate of the posterior probability function of a model being true ("AIC vs. BIC", 2007). The best model is the model with the lowest value.

Data analysis and ordinary linear regression were performed in R. To create the full model, the *lm* function was used, which fits linear regression models. Additionally, R helped analyze the

data by providing tools to select variables from the full model that should be included to create more significant models of predicting the amount of physical activity, and then analyzing these models based on the selection criterion mentioned above. Some of the tools used were forward, backward and stepwise selection, as well as the leaps package.

The leaps package performs an exhaustive search, where it compares all combinations given the number of covariates. The leaps package is beneficial because the algorithm returns the best model of each size, so the results do not depend on a penalty model for model size. Within the leaps package, a function called *regsubsets* helps pick out significant models. Based on these models, the selection criterion mentioned above were examined to decide on a good model of predicting the amount of physical activity.

Other approaches used for selecting models with significant variables from the full model were forward, backward and stepwise selection, using the *step* function. This function selects models based on their AIC value. In this study, only backward and stepwise produced models with fewer variables than the full model.

The overall predictive model chosen was based off of all the selection criterion mentioned earlier and the covariates included in each model. Two models were selected for comparison in the overall decision of the final model. By comparing adjusted R-square, residual sum of squares, R-square, Mallow's Cp, BIC, and AIC values, as well as looking at the significance of each covariate, and final model was chosen.

**Results**

To begin, the full model, which included all of the variables mentioned in the data description section, was analyzed. The *regsubsets* function provided eight models (one model with one covariate, one model with two covariates, etc.) with the most significant variables. Results from this function are shown in Table 1. As expected, *obesity* was included in all eight models. Several covariates were not included in any of the eight models. These include *UR*, *snap*, *internet*, *restaurants* and *sun*. Other variables that were included in five or more of the models were *temp*, *refund*, *uninsured* and *poverty*. These variables, as well as *obesity*, ended up being the best overall model for predicting the amount of physical activity, which will be discussed later.

For further analysis on these eight models, certain statistics such as adjusted R-square, residual sum of squares, R-square, Mallow's Cp, and BIC were examined. Models that had ideal values for these statistics were considered in the final analysis. Results of the covariates included in the best models based on the statistics mentioned above are shown in Table 2. The model selected from the *regsubsets* function with four covariates was considered the best model based on Cp and BIC. The models with six, seven and eight covariates were considered the best models based on adjusted R-square, R-square and residual sum of squares, respectively.

Additionally, backward and stepwise selection was performed on the full model. Backward and stepwise selection are based off of AIC values. For both backward and stepwise selection, the best model included *temp*, *refund*, *uninsured*, *poverty* and *obesity*. This model is the same model that was selected by the r*egsubsets* function for the best model with five covariates.

Located in Table 1 and Table 2, two models were examined further. Model one (1) and model two (2) are shown below.

$$Exercise = Obesity + Refund + Uninsured + Poverty + Intercept \quad (1)$$

$$Exercise = Obesity + Refund + Uninsured + Poverty + Temp + Intercept \quad (2)$$

The estimates and p-values for model (1) and model (2) are shown in Table 3 and Table 4, respectively. For model (1), every covariate was significant at the five percent confidence level. For model (2), every covariate except for *temp* was significant at the five percent confidence level.

Model (1) showed some interesting results. Shown in Table 3, the amount of physical activity is expected to decrease by approximately two units for every standard deviation increase in the amount of tax refund. For every unit increase in the poverty rate, the amount of physical activity is expected to decrease by .76. For every unit increase in *obesity*, the amount of physical activity is expected to decrease by .65 units, and for every unit increase in *uninsured*, the amount of physical activity is expected to increase by .47 units.

Model (2) showed similar results to model (1). Similar to model (1), the amount of physical activity is expected to increase for every unit increase in *uninsured*, to decrease for every unit increase in *poverty*, and *obesity,* and to decrease for every increase in standard deviation for *refund*. However for *uninsured*, the amount of physical is expected to increase less than in model (1), and for *poverty, obesity* and *refund*, the amount of physical activity is expected to

decrease less than in model (1). Furthermore, in model (2), the amount of physical activity is expected to decrease for every unit increase in *temp*.

To decide on the best model between model (1) and model (2), a comparison of certain statistics, shown in Table 5, was executed. Model (1) was considered the best model based on BIC and Cp values, but model (2) was considered the best model based on adjusted R-square, R-square, residual sum of squares, and AIC values. However, the differences in these values for model (1) and model (2) are very small. Therefore, both models are very similar in their ability to accurately predict the amount of physical activity. Additionally, *temp* was not significant at the five percent confidence level when included with *obesity, refund, uninsured* and *poverty*. Based on the analysis used in this study, the best model for predicting the amount of physical activity throughout the United States is model (1).

**Discussion**

There are several variables that were intended to be included in this study, but data was not available for every state in the United States. These variables include alcohol use, average number of hours of sleep each night, average number of hours spent at work every day, average amount of time watching television daily, and amount of money spent eating out at restaurants. These variables would have contributed further discussion to the study. This could be a possible area for further analysis if data was available.

Throughout the study, several covariates were never included in any of the models. These variables were *UR, snap, internet, restaurants* and *sun*. This was a surprising result, because it was expected that they somewhat contribute to the amount of physical activity performed across

the United States. This result might have occurred because of the accuracy of the dataset. The unemployment rate and snap participation were relatively easy to obtain accurate data for, but with the number of restaurants, average hours of sun and percent of individuals with internet connection on multiple devices there is room for error in the data.

For both model (1) and model (2), the amount of physical activity decreased for every increase in *tax, poverty* and *obesity*, but increased for every increase in *uninsured*. A decrease resulting from an increase in *poverty* is expected because as the average annual poverty rate increases, the less likely adults are going to be exercising, because they may not be able to afford it. Two results that weren't expected were how the amount of physical activity is expected to change as *refund* and *uninsured* increase. If tax refund increases, one would expect physical activity to increase because there is more money to spend on gym memberships, but instead the amount of physical activity decreased. Additionally, if the percent of uninsured individuals increased, it would be expected that less people would exercise because of the fear of getting hurt and not having insurance to cover the costs. Furthermore, the amount of physical activity due to *obesity* was somewhat expected because as obesity increases, the amount of physical activity one is able to engage in decreases. On the other hand, as obesity increases, the motivation to exercise more might increase.

Another result that was not expected was for forward selection not to eliminate covariates and not create a model other than the full model. Additionally, backward and stepwise selection resulted in the same model.

There are various options for further analysis. One possible extension for this study is to examine states individually by breaking them down into more in-depth analysis. Another extension could be to use another type of regression, such as nonparametric regression. Nonparametric regression could be used to guard against outliers, instead of scaling the variables to adjust the range. Lastly, some sort of transformation, such as a log-transformation could have been applied to the model in order to more accurately predict the amount of physical activity performed.

**References**


Centers for Disease Control and Prevention. (2010). *State Indicator Report on Physical Activity, 2010*. Retrieved from http://www.cdc.gov/physicalactivity/downloads/PA_State_Indicator_Report_2010.pdf

Centers for Disease Control and Prevention. (2013). *Obesity Prevalence Maps.* Retrieved from http://www.cdc.gov/obesity/data/table-adults.html

Centers for Disease Control and Prevention (2014). *Facts about Physical Activity.* Retrieved from http://www.cdc.gov/physicalactivity/data/facts.html

Centers for Disease Control and Prevention. (2008). *How much physical activity do adults need?* Retrieved from http://www.cdc.gov/physicalactivity/everyone/guidelines/adults.html

Centers for Disease Control and Prevention. (2011). *Physical Activity and Health.* Retrieved from http://www.cdc.gov/physicalactivity/everyone/health/index.html

Centers for Disease Control and Prevention. (2014). A*dult Obesity Facts*. Retrieved from http://www.cdc.gov/obesity/data/adult.html

Credit Karma. (2015). [Map illustration of the average credit score by state in the United States]. *Credit Trends.* Retrieved from https://www.creditkarma.com/trends/state

Current Results. (2015). *Average Annual Temperature for Each US State.* Retrieved from http://www.currentresults.com/Weather/US/average-annual-state-temperatures.php

Current Results. (2015). *Average Annual Sunshine by State.* Retrieved from http://www.currentresults.com/Weather/US/average-annual-state-sunshine.php

Governing. (2011). *Average Tax Refund by State.* Retrieved from http://www.governing.com/gov-data/average-irs-tax-refund.html

Governing. (2012). *State Poverty Rates, Averages.* Retrieved from http://www.governing.com/gov-data/other/state-poverty-rates-official-supplemental-poverty-measure.html

Governing. (2012). [Bar graph illustration of the state SNAP participation rates]. *Who Is On Food Stamps, By State.* Retrieved from http://www.governing.com/gov-data/food-stamp-snap-benefits-enrollment-participation-totals-map.html

Governing. (2011). *Internet Connectivity, Usage Statistics for States.* Retrieved from http://www.governing.com/gov-data/internet-usage-by-state.html

Max Mind. (2014). *Average Latitude and Longitude for US States.* Retrieved from http://dev.maxmind.com/geoip/legacy/codes/state_latlon/

The Methodology Center. (2007). *AIC vs. BIC.* Retrieved from http://methodology.psu.edu/eresources/ask/sp07

Noss, A. (September, 2014). *Household Income: 2013*. Retrieved from https://www.census.gov/content/dam/Census/library/publications/2014/acs/acsbr13-02.pdf

Smith, J.C., Medalia, C. (September, 2014). *Health Insurance Coverage in the United States: 2013.* Retrieved from https://www.census.gov/content/dam/Census/library/publications/2014/demo/p60-250.pdf

State Master. (2015). *Restaurants by state.* Retrieved from http://www.statemaster.com/graph/lif_res-lifestyle-resturants

U.S. Bureau of Labor Statistics. (2013). *Unemployment Rates for States*. Retrieved from http://www.bls.gov/lau/lastrk13.htm

**Appendix I**

**Table 1.** Summary of significant variables from *regsubsets* function.

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| **Unemployment Rate** | | | | | | | | |
| **Longitude** | | | | | | | | X |
| **Latitude** | | | X | | | | | |
| **Income** | | | | | | X | X | X |
| **Temperature** | | X | | | X | X | X | X |
| **Hours of Sun** | | | | | | | | |
| **Refund** | | | X | X | X | X | X | X |
| **Credit Score** | | | | | | | X | X |
| **Uninsured** | | | | X | X | X | X | X |
| **Poverty Rate** | | | | X | X | X | X | X |
| **Snap** | | | | | | | | |
| **Internet** | | | | | | | | |
| **Restaurants** | | | | | | | | |
| **Obesity** | X | X | X | X | X | X | X | X |

**Table 2.** Summary of significant variables based on certain statistics.

| | RSS | $R^2$ | Adj $R^2$ | Cp | BIC | AIC (stepwise) | AIC (backward) |
|---|---|---|---|---|---|---|---|
| **Unemployment Rate** | | | | | | | |
| **Longitude** | X | | | | | | |
| **Latitude** | | | | | | | |
| **Income** | X | X | X | | | | |
| **Temperature** | X | X | X | | | X | X |
| **Hours of Sun** | | | | | | | |
| **Refund** | X | X | X | X | X | X | X |
| **Credit Score** | X | X | | | | | |
| **Uninsured** | X | X | X | X | X | X | X |
| **Poverty Rate** | X | X | X | X | X | X | X |
| **Snap** | | | | | | | |
| **Internet** | | | | | | | |
| **Restaurants** | | | | | | | |
| **Obesity** | X | X | X | X | X | X | X |

**Table 3.** Summary of estimates and p-values for model (1).

| | Estimate | P-value |
|---|---|---|
| **Intercept** | 88.8809 | < 2 e-16 |
| **Refund** | - 2.0126 | 2.54 e-05 |
| **Uninsured** | .4670 | .00191 |
| **Poverty Rate** | - .7591 | 8.77 e-5 |
| **Obesity** | - .6519 | 1.56 e-5 |

**Table 4.** Summary of estimates and p-values for model (2).

| | **Estimate** | **P-value** |
|---|---|---|
| **Intercept** | 87.0450 | < 2 e-16 |
| **Temperature** | - .7467 | .164234 |
| **Refund** | - 1.7639 | .000389 |
| **Uninsured** | .4288 | .004359 |
| **Poverty Rate** | - .6037 | .005334 |
| **Obesity** | - .6477 | 1.52 e-5 |

**Table 5.** Comparison of model (1) and model (2) based on certain statistics.

| | **RSS** | **$R^2$** | **Adj $R^2$** | **Cp** | **BIC** | **AIC** |
|---|---|---|---|---|---|---|
| **Model 1** | 342.7944 | .662 | .632 | .652 | -34.642 | 106.25 |
| **Model 2** | 327.8832 | .676 | .640 | .883 | -32.954 | 106.03 |

**Appendix II**

```
setwd("~/UMD/Capstone Project")
mydata = read.csv("dataset.csv")
#install.packages("leaps")

#scaling variables
mydata$longitude=scale(mydata$longitude)
mydata$latitude=scale(mydata$latitude)
mydata$income=scale(mydata$income)
mydata$temp=scale(mydata$temp)
mydata$sun=scale(mydata$sun)
mydata$refund=scale(mydata$refund)
mydata$credit=scale(mydata$credit)
mydata$restaurants=scale(mydata$restaurants)
fullmodel = lm(exercise ~ UR + longitude + latitude + income + temp + sun + refund + credit +
uninsured + poverty + snap + internet + restaurants + obesity, data = mydata)

#leaps package
library(leaps)
models=regsubsets(exercise ~ UR + longitude + latitude + income + temp + sun + refund +
credit + uninsured + poverty + snap + internet + restaurants + obesity, nbest= 1, data=mydata)
summary(models)
summary(models)$adjr2
summary(models)$rss
summary(models)$rsq
summary(models)$cp
summary(models)$bic
str(summary(models))

#best model based on AIC value with stepwise selection
stepbest = step(fullmodel, direction = "both")
summary(stepbest)

#best model based on AIC value with forward selection
forbest = step(fullmodel, direction = "forward")
summary(forbest)

#best model based on AIC value with backward selection
backbest = step(fullmodel, direction = "backward")
summary(backbest)

#based on factors, check models
model1 = lm(exercise~refund + uninsured + poverty + obesity, data=mydata)
model2 = lm(exercise~refund + uninsured + poverty + obesity + temp, data=mydata)
summary(model1)
```

```
summary(model1)$r.squared
summary(model1)$adj.r.squared
summary(model1)
summary(model2)
```