

Regression Analysis of Tracking Steps

Xuan Shi, Neil Mecham, Kun Liu

February 12, 2014

Applied Regression Analysis  
STAT 6450  
Spring 2014

**Abstract**

With the arrival of new technology, new tools are created to help people with activities that could be enhanced with the use of technology. A prime example of this is wristbands, which keep track of one’s daily exercise. To test whether these wristbands assist people who engage in physical activity, a study was conducted using graphic methods and simple linear regression models to investigate the effects of the device. Relationships between different types of physical activity that the device measures were studied, and models were used to predict the values of missing data and to estimate other values that may occur. Using statistical methods, it was shown in this study that in the case of one particular subject, a wristband did positively affect the results of the man’s physical activity levels.

**Introduction**

There has recently been an increase in the use of wristbands that track daily physical activity, such as steps taken and calories burned. Compared to older step counters, these trackers can convert the real time steps into calories burned based on different conversion factors with casual walking, power walking, or running (Banas, 2014). Some popular models include: Basis B1, Fitbit Flex, Nike+FuelBand, Up by Jawbone and Polar RCX5, with retail price from $ 100 to $ 470 (Hill, L. 2013).These devices send the information to the users so they may analyze the amount of physical activity they do each day. It is advertised that these devices encourage people to become more active by providing the user with numeric data that measures their daily physical activity; this could motivate those who do not normally engage in physical activity to become more active because they can evaluate their progress and get excited about improvement from a normally inactive lifestyle.

To test this theory, we collected data from a “Force Delta” wristband used by Mr. Z, the subject of our study. The wristband measured the following daily physical activity: steps taken, distance traveled (miles), active minutes, floors, and calories burned. Mr. Z used the device for 23 days before it unfortunately broke. We have taken the data and analyzed it using statistical methods to look for significant trends over time and the different measurements of daily physical activity.

The purpose of this analysis was to determine if there was a significant increase in Mr. Z’s physical activity over the 23 days he wore the Force Delta wristband. This would support the hypothesis that wristbands motivate people to become more physically active by quantifying their daily physical activity. Another goal is to find relationships between parameters recorded by the tracker to help us understand the functions of the tracker and make predictions for some missing data.

**Methods**

Graphs plotting the relationship between each pair of variables were created and examined to provide a way of briefly examining the data (Figure 1).Since the purpose of this study is to determine whether the device helps people pursue a healthier lifestyle, the “Weekday” variable in the original dataset was replaced by the “Order” variable, that is the observational order of the recordings. These plots allowed us to look for trends or outliers in the data that we might wish to investigate further.

By plotting the relationship between each pair of variables, we did find one significant outlier on Day 23.The calories count for this day was 999999, but all other observations for that day were reasonable. The subject reported that the device malfunctioned on the 23rd day, so we are convinced that the extreme observation is caused by the malfunctioning of the tracker. As a result, we decided to discard this piece of data but wished to include day 23 in our observations. To provide an estimate for the true value of calories burned on that day, we created a regression model between calories burned and each other variable (steps, floors, distance, active minutes) and chose the regression model with the variable that had the closest linear relationship to calories burned. We then replaced the calorie count of Day 23 with estimation from the best regression model. The process of this is discussed further in this paper.



Figure1. Relationship between variables of all data

Another potential problem was when we were told by Mr. Z that he had forgotten to wear his Force Delta wristband a couple of days during the 23 day period, but he had forgotten which days. We examined the data and deleted observations 6 and 7 from our data based on the fact that both days had abnormal step measurements compared to other observations (dramatically lower than normal). A possible explanation of the oddities in the data is that Mr. Z forgot to wear it during most the day and wore it later in the afternoon or evening. After eliminating data of these two days, the full analysis was conducted with this new set of data.

With the corrected dataset, we examined trends in the data using graphical procedures. Regression models were created, and linear tests were run on these models to test the hypothesis that physical activity increased over the 23 day period.



Figure2. Relationship between variables with corrected data

**Construction of Model for Calories Burned**

The graphs above indicate a pattern of linear associations between variables. As a result, we construct a simple linear regression model as follows:

 (1)

Where:

is the value of the calories burned on the th day

and are parameters

 is a known constant and the value of the predictor variable on the th day

 is a random error term with mean  and ; and  are uncorrelated for all ,,



In this particular case, to set up interval estimates and make tests, we adopt the assumption that the error terms  are normally distributed. Thus, our model turns into the normal error regression model, as follows:

 (2)

Where:

 is the value of the calories burned on the th day

 and  are parameters

 is a known constant and the value of the predictor variable on the th day

 are independent 



Figure 2 shows the relationship between each pair of variables excluding outlier and abnormal observations. From the plot, we can find that steps, distance, floors and active minutes all show some linearity with calories. To make best prediction of calories, we need to find the best-fit regression model based on these four variables.

The working mechanism of wristband excludes “Floors” and “Active Minutes” as predictor variables to predict the Calories burned. This is because either of them measures a special case of exercise, and cannot reflect the overall situation. For example, the “Floors” counts the floors one takes daily, but the subject may live and do exercise on the first floor. As a result, it fails to reflect the true level of exercise and the calories burned. Similarly, “Active Minutes” counts the time when the subject is doing some “active” sports or exercise(such as running), yet it doesn’t count the time one is doing “less active” exercise such as walking. Further comparisons concerning the coefficient of determination (R2) for each pair indicates that steps has the best prediction capability among the four variables with R2= 0.924 (Figure 3). Thus, we determine to regress calories on steps, and obtain the regression function:

 (3)

Where:

 is the fitted value of the calories burned on the th day

 and  are maximum likelihood estimators of parameters  and 

 is the value of steps on the th day and a known constant

Calculation from the modified dataset yields that,

 (4)

Thus, the fitted regression function is:

 (5)

and residual is:

 (6)

**Checking aptness of Model**

1. Tests concerning 

To convince researchers of the linear relationship between calories and steps, a two-sided test whether or not  is taken. Namely,

 (7)

The test statistic:

 (8)

is calculated(where  is the estimated variance of ) and . The P-Value of the test is less than 0.0001, so the test is to reject  and conclude. That is, the slope of the fitted regression function is not 0. This is the same as the plot in Figure 3 shows.



Figure3. Regression function based on steps

Further investigation leads to the conclusion that the coefficient of correlation, which indicates strong, positive correlation between calories burned and steps.

2. Diagnostics for Residuals

The independence assumption of the error term is examined in Figure 4. As we expected, the residuals bounce randomly around the residual = 0, which suggests independence.



Figure4. Residual independence check

The assumption of constant variance of the error terms is examined in Figure 5 and 6. Despite the fact that the sample size of 20 may not be sufficiently large, both the residuals and the absolute residuals plotted against predictor variable showed no systematic trend to vary. As a result, these plots support the assumption of constant variance of the error terms.

Figure6. Absolute Residuals versus Predictor Variable

Figure5. Residuals versus Predictor Variable

The normality assumption is checked in Figure 7 by the Distribution Plot, where residuals are plotted against the expected value under Normality. Under Normality, the expected value is computed by the formula:

 (9)

Where

is the rank of the th residual in the pooled residuals

is the  percentile of the standard normal distribution

The plot shows a nearly linear pattern, which suggests agreement with normality assumption.



Figure7. Normal probability plot for residual

A further correlation test for normality is conducted by calculating the coefficient of correlation between residuals and the expected values under normality:

 (10)

where

is the true value of residual

is the expected value of residual under normality

Calculation yields that, providing a strong support for the normality assumption as the Normal Distribution Plot indicates.

The analyses and diagnostics above indicate that our model is appropriate with the assumptions satisfied.

**Testing for Increase in Activity**

It is difficult to measure an increase in activity level since there are so many different ways of quantifying activity level, such as steps taken, calories burned, and active minutes (all of which the tracker measures).We decided that for our purposes, an increase in steps taken most accurately reflected an increase in activity level, because the purpose of the tracker is not necessarily to start exercising everyday (like active minutes), but to encourage people to spend more time every day being physically active, and walking is the most common physical activity that is measured by the tracker.

To check whether trackers motivate people to do better in terms of an increase in activity, we decided to test whether the recordings of steps taken show a pattern to increase over time. Although a formal test would be optimal, it's not proper to use the observation order as a predictor variable in this regression model. This is due to the fact that observation order and the step recordings are not related by a linear association. This was discovered when we found the coefficient of determination between steps and observation order to be 0.1237, indicating a poor linear association. As a result, the test for increase in activity is conducted by analyzing the graphs plotting the different recordings taken by the tracker over observation order.

**Results and Discussion**

With the first regression model produced, we may give a good prediction of calories on day 23 (Figure 8). The prediction mean is 2597, with 95% confidence interval [2363, 2831].



Figure 8. New prediction for day 23

Based on the model, the y intercept is 1755, which may be a prediction for calories burned in a day that Mr. Z is sedentary, such as sitting in a couch for a whole day.

However, the selected observation minimum is 5157, which is not very close to zero, so the accuracy may not be exceptional. At least the intercept gives us a reference about the calorie starting point of Mr. Z’s tracker, which may also vary between different people. According to the Healthy Eating Guide, sitting burns an estimated 75 calories per hour (Meara, A. 2013). The number of actual calories burned depends on body composition, weight and metabolism.

For the tested 20 days, the change of recordings is plotted (Figure 9). From these plots, we see an increasing trend, although not obvious due to the relatively small sample size, with several fluctuations. But we can see after day 10, Mr. Z’s recordings are generally higher than those before day 10. This may indicate the gradual change of his lifestyle. Those fluctuations are also reasonable as Mr. Z just started to wear this tracker, he may not yet fully adapted to the change to his life. Longer observation and monitoring is required to convincingly conclude that the step track or wrist band can really make a difference in getting people more involved in physical activities.



Figure 9. Observed recordings over 20 days

**Conclusion:**

We have studied the relationships between the different measurements and have found a strong linear relationship between steps taken and calories burned. This allowed us to provide an estimate of calories burned on Day 23 and proceed with our analysis of the data using this estimate. This model also allows us to make estimations and predictions on the number of calories burned based on the number of steps taken even without the use of the tracker. This means that we can provide an estimate for days when Mr. Z does nothing but sit on the couch (although estimates generally become more accurate as the number of steps increases).This also means that Mr. Z could estimate the number of calories he burned each day with a simpler step counter than the Delta Force, although it would require a bit more work on his part.

We also concluded that the Delta Force motivated Mr. Z to increase his physical activity over the 23 day period. More observations and monitoring are required to convincingly conclude that a step tracker or wrist band can really make a difference in getting the general population more physically active. However, we know that it worked in this particular case over a period lasting about a month. Because of this, we do believe that a step tracker on the wrist or in the pocket could possibly help motivate people to take more steps and become more physically active than they would otherwise be.

**References:**

1. Kutner, M., C. Nachtsheim, J. Neter. (2004). Applied Linear Regression Models: 4th edition. McGraw-Hill/Irwin.
2. Banas, T. (2014). How to convert pedometer steps to calories. livestrong.com. Retrieved Feb. 10, 2014, from http://www.livestrong.com/
3. Hill, L. (2013, June 27). The Best Fitness Tracker Bracelets. businessweek.com. Retrieved Feb. 10, 2014, from http://www.businessweek.com/
4. Meara, A. (2013, August 16). How many calories can you burn daily just from sitting & walking around a house? Livestrong.com. Retrieved Feb. 10, 2014, from http://www.livestrong.com/

Appendix (Original R code):

## Load librarys ##

library(MASS)

library(graphics)

## original dataset ##

data.exercise = read.csv("Force-Delta-log.csv",header = T, sep = ",");

data.exercise = data.frame(data.exercise)

attach(data.exercise)

## modified dataset-eliminate outlies and wrong data ##

data.tracker = read.csv("proj1.csv",header = T, sep = ",");

data.tracker = data.frame(data.tracker)

attach(data.tracker)

## ------- Simple Plotting --------------------##

pairs(data.exercise)

pairs(data.tracker)

## ------Regress Calories on 4 seperate variables------------##

Y=data.tracker$tCalories

X.steps=data.tracker$tSteps

X.distance=data.tracker$tDistance

X.floors=data.tracker$tFloors

X.actminute=data.tracker$tActiveMinutes

## Regress on steps##

b.1.steps=sum((X.steps-mean(X.steps))\*(Y-mean(Y)))/sum((X.steps-mean(X.steps))^2)

b.0.steps=mean(Y)-b.1.steps\*mean(X.steps)

Y.hat.steps=b.0.steps + b.1.steps\*X.steps

e.steps=Y-Y.hat.steps

## Regress on distance ##

b.1.distance=sum((X.distance-mean(X.distance))\*(Y-mean(Y)))/sum((X.distance-mean(X.distance))^2)

b.0.distance=mean(Y)-b.1.distance\*mean(X.distance)

Y.hat.distance=b.0.distance + b.1.distance\*X.distance

e.distance=Y-Y.hat.distance

## Regress on floors ##

b.1.floors=sum((X.floors-mean(X.floors))\*(Y-mean(Y)))/sum((X.floors-mean(X.floors))^2)

b.0.floors=mean(Y)-b.1.floors\*mean(X.floors)

Y.hat.floors=b.0.floors + b.1.floors\*X.floors

e.floors=Y-Y.hat.floors

## Regress on active minutes ##

b.1.actminute=sum((X.actminute-mean(X.actminute))\*(Y-mean(Y)))/sum((X.actminute-mean(X.actminute))^2)

b.0.actminute=mean(Y)-b.1.actminute\*mean(X.actminute)

Y.hat.actminute=b.0.actminute + b.1.actminute\*X.actminute

e.actminute=Y-Y.hat.actminute

## --------Check the best-fit regression model----------##

SSTO=sum((Y-mean(Y))^2)

n=20

SSE.steps=sum(e.steps^2)

MSE.steps=SSE.steps/(n-2)

SSE.distance=sum(e.distance^2)

SSE.floors=sum(e.floors^2)

SSE.actminute=sum(e.actminute^2)

Rsquare.steps = (SSTO-SSE.steps)/SSTO

Rsquare.distance = (SSTO-SSE.distance)/SSTO

Rsquare.floors = (SSTO-SSE.floors)/SSTO

Rsquare.actminute = (SSTO-SSE.actminute)/SSTO

Rsquare.steps

Rsquare.distance

Rsquare.floors

Rsquare.actminute

##--------Test Hypothesis if Beta.1.Steps =0-----#

s.b1.steps=sqrt(MSE.steps/sum((X.steps-mean(X.steps))^2))

t.star.steps=b.1.steps/s.b1.steps

t.star.steps

qt(0.975,n-2)

##----------Use Regression Model on steps -----##

X.steps.new = 12765

Y.hat.new =b.0.steps + b.1.steps\*X.steps.new

alpha = 0.05

t=qt(1-(alpha/2),n-2)

plot(data.tracker$tSteps,data.tracker$tCalories,cex=1.5,cex.lab=1.5,xlab="Steps",ylab="Calories Burned")

abline(b.0.steps, b.1.steps,col=2,lwd=3)

##---------Prediction Interval for 23rd day--------##

Nume=(X.steps.new-mean(X.steps))^2

Deno=sum((X.steps-mean(X.steps))^2)

add=1+(1/n)+(Nume/Deno)

s.square.pred = MSE.steps\*add

s.pred = sqrt(s.square.pred)

e.steps.star=e.steps/sqrt(MSE.steps)

lower = Y.hat.new-t\*s.pred

upper = Y.hat.new+t\*s.pred

Y.hat.new

lower

upper

##------plot the 95 percent Prediction Interval with Predicted value----##

##-can change confidence by setting alpha above another value-##

plot(data.tracker$tSteps,data.tracker$tCalories,cex=1.5,cex.lab=1.5,xlab="Steps",ylab="Calories Burned")

abline(b.0.steps, b.1.steps,col=2,lwd=3)

lines(X.steps.new,Y.hat.new,type="p",col=4,lwd=2)

segments(X.steps.new,lower,X.steps.new,upper,col=4,lwd=2)

##-----------Check Model's Assumptions----------##

##-------Check independent epsilons----##

plot(seq(1:20),e.steps,cex=1.5,cex.lab=1.5,xlab="Observation Order",ylab="Residuals",type="l",lwd=1.5)

lines(e.steps,type="p")

abline(0,0,col=2,lwd=2)

##-----Check Constant Variance of epsilons---##

plot(X.steps, e.steps,cex=1.5,cex.lab=1.5,xlab="Steps",ylab="Residuals",lwd=1.5)

e.steps.absolute=sqrt(e.steps^2)

plot(X.steps, e.steps.absolute,cex=1.5,cex.lab=1.5,xlab="Steps",ylab="Absolute Residuals",lwd=1.5)

##-----Check for Normality-------##

rank=rank(e.steps)

expected.value = sqrt(MSE.steps)\*qnorm((rank-0.375)/(n+0.25))

plot(expected.value,e.steps,cex=1.5,type="p",cex.lab=1.5,xlab="Expected Value under Normality",ylab="Residual")

abline(0,1,col=2,lwd=2)

rnume=sum((e.steps-mean(e.steps))\*(expect-mean(expect)))

rdeno=sqrt(sum((e.steps-mean(e.steps))^2)\*sum((expect-mean(expect))^2))

r=rnume/rdeno

##-----Plots to see a healthier lifestyle--##

plot(seq(1:20),tSteps,type="p",cex=1.5,xlab="Observation Order",ylab="Steps",cex.lab=1.5)

plot(seq(1:20),tDistance,type="p",cex=1.5,xlab="Observation Order",ylab="Distance",cex.lab=1.5)

plot(seq(1:20),tFloors,type="p",cex=1.5,xlab="Observation Order",ylab="Floors",cex.lab=1.5)

plot(seq(1:20),tActiveMinutes,type="p",cex=1.5,xlab="Observation Order",ylab="ActiveMinutes",cex.lab=1.5)

plot(seq(1:20),tCalories,type="p",cex=1.5,xlab="Observation Order",ylab="Calories",cex.lab=1.5)