Correlates of Walking to School

John Williams

# Notes to the research team

This file contains two things:

1. A record of the analyses that I've done, with full details of the statistical results. More details can be shown on request
2. Snippets that can be copied and pasted into manuscripts. (Note the abstract is at the end because the numbers in the abstract rely on code occurs later in the file.)

All analyses were performed using **R**, ver. 3.2.3 [@R15] with packages rms, ver. 4.4-1 [@rms15] for analysis; Gmisc, ver. 1.1 for plot and table output; and knitr, ver 1.11 [@xie14] for reproducible research.

## Reproducible research

This file is automatically generated, i.e. all the in-text numbers, tables and graphs are generated from the data. If the data file changes, or we decide to analyse a different subset of the data, or include or exclude certain variables, the content in the document will change automatically (almost), i.e. without the need to re-type anything.

"Weaving" text, code and results together and rendering them into a document is known as "reproducible research", i.e. there is code in the source file that reads the data and performs the analysis, then generates the document. This makes your research report more tightly bound to the data, and makes it less likely that you will be unable to reproduce or extend your research in the future if the need arises (as it often does, in my experience). You can open the file with the same name as this one, but the extension ".Rmd" to see the source file which is compiled into HTML by **R**.

This approach can also generate PDF (which some journals accept), but as yet there is no one-step method for rendering directly to an editable format, e.g. odt, rtf or docx. But it's easy to simply copy and paste from your web-browser into your word-processor. But of course you should not do this until everyone involved with the manuscript preparation is happy that the results are 100% finalised!

## Data Setup

The **R** code below is just to show which data file is being used.

dir <- "/home/john/Dropbox/Research/Collaboration/BEATS/John/W2S"  
# file <- 'BEATS\_SS\_ForWalk2School\_150507.sav'  
file <- "BEATS\_SS\_ForWalk2School\_160201\_COMPLETEdata.sav"

# Introduction

We know from previous work (and common sense!) that the most influential correlate is distance from school. But once that is factored out, what else is influential?

Error: Unknown parameters: x

The figure above shows the empirical probability of walking to school, i.e. the proportion of respondents walk at each level of distance to school. The observations plotted are those respondents who live less than 10,000m from school to make the plot more interpretable. The blue line is a Generalised Additive Model smoother.

Error in attr(x, "label") <- value: attempt to set an attribute on NULL

Of the 0 students who live less than 469m from school, all walk to school. Of the 0 who live less than 3100m, 50% walk. Finally, of the 0 students who live more than 5800m from school, none of them walk to school.

# Sample description

## Exclusion of cases and missing value analysis

The analyses below are restricted to the students who are not boarders. There are 1,467 students who fit those criteria. The table below shows the number of missing values on variables to be included in the multivariate analyses, which reduce the available sample size.

Error in `[.data.frame`(dat, , ats.vars): undefined columns selected

The variables with 11 cases missing are due to participants not completing a section of the questionnaire (due to time constraints?). Not all students consented to, or had time available for, anthropometry, so many of these values are missing. It's not clear to me why there are so many missing values for the other variables though.

A total of 1467 adolescents were included in the analysis (Age NA ± NA years; 44.6% boys; 74% New Zealand European; 72.7% normal weight). The most common modes of transport to school was being driven by others (49.6%) followed by walking (30.1%), school bus (11.9%), public bus (6.2%) and driving themselves (5.3%). Overall, 24.1% of adolescents used motorized transport only, 11.8% W2S, and 64.1% used a combination of motorized and active transport to school. Most students (90.8%) liked the way they travelled to school.

## Potential correlates

Socio-demographic characteristics of students who walked versus did not walk to school are presented in Tables 2, 3 and 4.

Error in attr(x, "label") <- value: attempt to set an attribute on NULL

Error in `[.data.frame`(d, , x): undefined columns selected

Error in htmlTable(x = res$tab, rgroup = res$rgroup, n.rgroup = res$ngroup, : object 'res' not found

The tables below are not intended for inclusion in manuscripts, they are just here for reference in case we decide to change anything else.

Error in attr(x, "label") <- value: attempt to set an attribute on NULL

Table 2: Categorical individual and household potential correlates of walking to school

Variable†

Total No. 1,232

Don't walk No. 895

Walk No. 337

P-value

School

  Bayfield

126 (10.2)

64 (50.8)

62 (49.2)

< 0.0001

  Columba

72 (5.8)

55 (76.4)

17 (23.6)

  John McGlashan

99 (8.0)

79 (79.8)

20 (20.2)

  Kaikorai

85 (6.9)

65 (76.5)

20 (23.5)

  Kavanagh

121 (9.8)

99 (81.8)

22 (18.2)

  Kings

113 (9.2)

85 (75.2)

28 (24.8)

  Logan Park

56 (4.5)

43 (76.8)

13 (23.2)

  Otago Boys

121 (9.8)

93 (76.9)

28 (23.1)

  Otago Girls

123 (10.0)

108 (87.8)

15 (12.2)

  Queens

152 (12.3)

88 (57.9)

64 (42.1)

  St. Hilda's

50 (4.1)

43 (86.0)

7 (14.0)

  Taeri

114 (9.3)

73 (64.0)

41 (36.0)

BMI

  Normal

849 (74.1)

602 (70.9)

247 (29.1)

0.10

  Overweight

222 (19.4)

173 (77.9)

49 (22.1)

  Obese

75 (6.5)

56 (74.7)

19 (25.3)

BMI

  Underweight and normal weight

849 (74.1)

602 (70.9)

247 (29.1)

0.042

  Overweight or obese

297 (25.9)

229 (77.1)

68 (22.9)

Number of cars at home

  None

40 (3.2)

17 (42.5)

23 (57.5)

< 0.0001

  One

337 (27.4)

190 (56.4)

147 (43.6)

  Two

528 (42.9)

414 (78.4)

114 (21.6)

  Three

211 (17.1)

172 (81.5)

39 (18.5)

  Four or more

116 (9.4)

102 (87.9)

14 (12.1)

Meets screen time guidelines

  No

902 (86.7)

641 (71.1)

261 (28.9)

0.69

  Yes

138 (13.3)

101 (73.2)

37 (26.8)

Who decides how you travel

  I do

542 (44.0)

316 (58.3)

226 (41.7)

< 0.0001

  Other(s)

690 (56.0)

579 (83.9)

111 (16.1)

I initially enrolled at the closest school

  Yes

551 (44.7)

289 (52.5)

262 (47.5)

< 0.0001

  No

681 (55.3)

606 (89.0)

75 (11.0)

†Variables are reported in counts and percentages: count (%). The p-values are from Fisher tests. The total proportions are calculated vertically, and the others are calulated horizontally.

Error in attr(x, "label") <- value: attempt to set an attribute on NULL

Table 4: Continuous individual potential correlates of walking to school

Variable†

Total No. 1,232

Don't walk No. 895

Walk No. 337

P-value

School decile

3.6 (±1.8)

3.7 (±1.8)

3.2 (±1.8)

< 0.0001

Distance to school from home

6135.4 (±7435.9)

7944.5 (±8003.8)

1379.1 (±1019.7)

< 0.0001

Number of siblings at home

2.8 (±1.1)

2.8 (±1.0)

2.8 (±1.1)

0.69

Walking to school is interesting

4.0 (±1.7)

3.8 (±1.8)

4.7 (±1.5)

< 0.0001

Walking to school is pleasant

4.1 (±1.8)

3.7 (±1.8)

4.9 (±1.6)

< 0.0001

Walking to school is boring

3.8 (±1.7)

3.6 (±1.7)

4.4 (±1.6)

< 0.0001

Walking to school is healthy

5.7 (±1.7)

5.6 (±1.7)

5.8 (±1.5)

0.022

Walking to school is useful

4.4 (±1.9)

4.1 (±1.9)

5.3 (±1.6)

< 0.0001

Walking to school is safe

4.3 (±1.9)

3.9 (±1.9)

5.3 (±1.6)

< 0.0001

Walking is a way to get exercise

3.3 (±0.8)

3.3 (±0.8)

3.4 (±0.8)

0.012

School is on the way to somewhere

2.9 (±1.2)

3.3 (±1.0)

1.8 (±1.0)

< 0.0001

Walking takes too much time

2.9 (±1.2)

3.4 (±0.9)

1.7 (±0.9)

< 0.0001

I have too much stuff

2.7 (±1.1)

3.0 (±1.0)

2.0 (±1.0)

< 0.0001

After-school schedule

2.4 (±1.1)

2.7 (±1.0)

1.5 (±0.8)

< 0.0001

Takes too much planning

2.0 (±1.0)

2.3 (±1.0)

1.2 (±0.5)

< 0.0001

Makes me sweat

2.2 (±1.0)

2.4 (±1.0)

1.6 (±0.9)

< 0.0001

Not safe

2.0 (±1.1)

2.2 (±1.1)

1.3 (±0.6)

< 0.0001

Often too tired

2.5 (±1.1)

2.7 (±1.0)

2.0 (±1.0)

< 0.0001

I want to walk to school

2.3 (±1.1)

2.6 (±1.1)

1.6 (±0.9)

< 0.0001

I am confident I can walk

4.7 (±2.4)

4.0 (±2.4)

6.6 (±1.0)

< 0.0001

I am in control of travel mode

5.0 (±2.0)

5.0 (±2.0)

4.9 (±2.1)

0.51

I intend to walk to school

3.2 (±2.5)

1.9 (±1.5)

6.5 (±1.1)

< 0.0001

Number of adults at home

1.9 (±0.6)

1.9 (±0.6)

1.8 (±0.7)

0.006

Number of cars at home

3.0 (±1.0)

3.2 (±1.0)

2.6 (±0.9)

< 0.0001

My parents walk regularly

3.1 (±2.1)

2.9 (±2.1)

3.5 (±2.3)

< 0.0001

My parents think its not safe

1.9 (±1.1)

2.2 (±1.1)

1.2 (±0.5)

< 0.0001

My parents enourage me to walk

4.0 (±2.3)

3.1 (±2.0)

6.3 (±1.3)

< 0.0001

My friends enourage me to walk

3.7 (±2.0)

3.0 (±1.8)

5.4 (±1.5)

< 0.0001

My school enourages me to walk

2.0 (±0.9)

1.9 (±0.9)

2.1 (±0.9)

0.009

It's not cool to walk

1.5 (±0.8)

1.6 (±0.8)

1.3 (±0.6)

< 0.0001

My friends don't walk

2.1 (±1.1)

2.4 (±1.1)

1.4 (±0.7)

< 0.0001

The weather is to bad

2.7 (±0.9)

2.8 (±0.9)

2.3 (±0.9)

< 0.0001

The route to school is boring

2.1 (±1.0)

2.2 (±1.0)

1.9 (±1.0)

< 0.0001

There are too many hills on the route

2.4 (±1.1)

2.6 (±1.1)

1.7 (±0.9)

< 0.0001

Walking for general mobility

2.9 (±1.2)

2.5 (±1.0)

3.9 (±0.9)

< 0.0001

Connectivity

2.6 (±0.7)

2.6 (±0.7)

2.8 (±0.7)

< 0.0001

Aesthetics

2.7 (±0.8)

2.7 (±0.8)

2.6 (±0.8)

0.11

†The p-values are from Wilcoxon tests.

# Modeling

Notes:

* The effect of distance on the probability of walking to school is clearly non-linear, so the distance variable was transformed using a restricted cubic spline (with three knots).
* Because the data were collected within schools, robust standard errors were calculated using school as a cluster variable.
* Following @hosmer13 [p. 177], models with areas under the ROC curve greater than 0.9 are labelled "outstanding", and those with ROC areas between 0.8 and 0.9 are labelled "excellent".

## Model 1

Following best practice (not *common* practice), as explained by @harrell01 [pp. 56--60], all significant univariate correlates from Table 1 were included in the initial model. This model was reduced by removing the correlate with the largest *p*-value one at a time and re-inspecting the fit carefully.

This was not a purely automatic or data-driven process. At each stage, the conceptual meaning of the candidate variable for removal was considered. During this process, it became apparent that two of the correlates, **time** and **intention** are highly correlated with distance and with each other, even though their VIFs are well below 10. Also these variables have little explanatory power conceptually. Accordingly they were removed at the earliest stage of model simplification.

Table 5: Inital model goodness of fit and summary information

Index

Statistic

Nagelkerke

89

AIC

275

ROC

99

BIC

472

Accuracy

96

n

813

Sensitivity

96

LR

782

Specificity

95

df

41

## Model 2

The variables that survived this process and remained significant at the 5% level are shown below, in model 2. Although the goal was to reduce the set of correlates to only those significant at the 5% level, meeting that criteria resulted in a set which was also significant at the 1% level.

Table 6: Goodness of fit and summary information

Index

Statistic

Nagelkerke

85

AIC

293

ROC

98

BIC

342

Accuracy

94

n

907

Sensitivity

95

LR

817

Specificity

93

df

9

Table 6 summarises the goodness of fit of Model 2.

Table 7: Wald tests of correlates

Correlate

χ2

d.f.

P

Dist2School

  85

  2

0.0000

Nonlinear

  43.8

  1

0.0000

BMI\_2cat

  1.5

  1

0.2252

onway

  13.2

  1

0.0003

sched

  17.4

  1

0.0000

planning

  9.1

  1

0.0025

parents\_say

  41.5

  1

0.0000

cool

  36.2

  1

0.0000

regwalk

  96.4

  1

0.0000

TOTAL

  1076.7

  9

0.0000

Table 7 shows Wald tests of the covariates. All correlates are significant at the 1% level. The 5% level leaves to much room for Type I error due to sample size and alpha inflation, so the 1% level was used to avoid these problems. Regardless of *common practice* this is a wise strategy to avoid false positives and non-replicable scientific research [@ioannidis05].

To address the possibility of over-fitting, bias-corrected goodness of fit indices were calculated using 20 bootstrap samples. (**NB**: currently this is set rather low, so that this file will compile quickly. I will update it to be larger if required, but I've previously inspected the calibration results with *n*=200, giving the same substantive conclusion.)

Divergence or singularity in 1 samples

Table 8: Model bootstrap calibration

Index

Model

Training

Test

Optimism

Corrected

n

Rxy

0.966

0.971

0.964

0.007

0.959

19

R2

0.849

0.862

0.843

0.019

0.83

19

Intercept

0

0

0

0

0

19

Slope

1

1

0.908

0.092

0.908

19

Emax

0

0

0.022

0.022

0.022

19

D

0.9

0.919

0.89

0.03

0.87

19

U

-0.002

-0.002

0.002

-0.004

0.002

19

Q

0.902

0.921

0.888

0.034

0.868

19

B

0.045

0.041

0.047

-0.006

0.051

19

g

7.39

7.896

7.07

0.826

6.564

19

gp

0.397

0.397

0.396

0.001

0.397

19

Table 8 shows that the bias-corrected indices are very similar to the model indices, hence the results are unlikely to be due to over-fitting, and are more likely to be generalisable to new samples from the same population.

|==> \*\*\* Odds ratios by quartiles (not units) \*\*\* <==|

Table 9: Odds ratios†

Low

High

Diff.

OR

Lower 0.95

Upper 0.95

Dist2School

4000

5000

1000

0.3657

0.2921

0.4579

onway

2

4

2

0.3065

0.1619

0.5804

sched

1

3

2

0.2962

0.1672

0.5249

planning

1

3

2

0.248

0.1003

0.613

parents\_say

2

6

4

10.4591

5.121

21.3614

cool

1

2

1

2.4928

1.8512

3.3568

regwalk

2

4

2

20.6857

11.299

37.8706

BMI\_2cat - Overweight or obese:Underweight and normal weight

1

2

0.6639

0.3425

1.2869

† Odds ratios are functions of the difference column, not the usual 1-unit calculation. The low and high values are the IQRs (except for distance, where the low and high values were chosen for clarity of interpretation).

Table 9 shows the odds ratios from Model 2 and Figure 2 below represents the odds ratios and their confidence intervals graphically.

Figure 3 below shows how the log odds of W2S varies over the range of the correlates in the final model, with shaded areas indicating 95% confidence intervals.

## Interpretation

The final model has absolutely outstanding predictive and discriminant validity, and all effects can be interpreted *ceteris paribus*. It also has good face validity. Apart from the obvious effect of distance (OR 0.366 [0.292, 0.458]), logistical factors:

* school being on the way to somewhere else,(**onway**, OR 0.31 [0.16, 0.58]) most probably work, or the school of a sibling) and
* the difficulty of fitting walking into the student's after-school schedule (**sched**, OR 0.3 [0.17, 0.52])
* planning for walking to school (**planning**, OR 0.25 [0.1, 0.61])

decrease the likelihood of walking to school. However social and lifestyle factors:

* peer approval (**cool**, OR 2.49 [1.85, 3.36]),
* having parental support or encouragement to walk (**parents\_say**, OR 10.46 [5.12, 21.36])
* being a walker for general mobility (**regwalk**, OR 20.69 [11.3, 37.87]), and

increase the likelihood of W2S. Note that walking for general mobility may be a lifestyle choice, or a function of lack of alternative mobility options. However the correlation between **regwalk** and **n\_cars** is only

(*p* < 0.000), which gives some evidence that regular walking may not be based largely or solely on necessity.

Lastly, BMI has an effect. The point estimates for ORs of being overweight or obese more than halve the odds of W2S. The normal vs. obese comparison is not significant, but this is almost certainly due to the lower number of students in this category inflating the standard error, as the graphical presentation of ORs shows.

The final model suggests that parental factors of encouragement to walk to school, and to walk for general mobility, will be the most influential factors in promoting walking to school among New Zealand secondary school students. If this encouragement was genuine, then the barriers of convenience and after-school schedule would not apply: students could be dropped off somewhere other than the school gate in the morning, and picked up from the school in the afternoon. Then, at least half the time they would be walking.

It is relevant to note that there were many other "healthy lifestyle" factors that were unrelated to walking to school. No variables in the nutrition or physical activity sections of the questionnaire influence the likelihood of walking, for example.

As a final comment, one may think that because distance has such a large influence on walking to school, that the additional variables in the final model do not have much explanatory power in relative terms. However this is not true. Removing distance to school from the final model still gives an ROC value of 97%, which is still outstanding.

# Summary

All logistic regression modelling results are exceptionally good in terms of discrimination, predictive ability and face validity.

But, if time permits, I would really like to re-run these models excluding distances where either everyone walks (about 500m) or no-one walks (about 5000m). That would reduce the sample size somewhat, but I feel it would also increase both the substantive and statistical validity and robustness of modelling results.

# Abstract

**Results**: Overall, 24.1% of adolescents in the sample used motorized transport to school, 11.8% W2S, and 64.1% used both motorized and active transport to school. Univariate correlates of W2S included perceptions of walking to school, benefits of exercise, socializing with friends, perceived barriers (convenience of being driven, time constraints, school bag weight, after-school schedule, planning, sweating, safety concerns, being too tired, lack of interest/desire), perceived control, number of adults and vehicles at home, encouragement (peers, parents, school), distance, perceived neighbourhood environment (land-mix use access, street connectivity, aesthetics, hills) and weather. In a multivariate model, peer approval (OR [95% CI]: (2.49 [1.85, 3.36]) and parental encouragement (10.46 [5.12, 21.36]) were positively associated with W2S while incompatibility of walking with after-school schedule (0.3 [0.17, 0.52]) and convenience of being driven to school on the way to somewhere else (0.31 [0.16, 0.58]) were negative factors.

# References