# The Statistics of Cycling

Matthew Arnold
Grey College

3rd March 2010

# Contents

- Data
- Usage Profiles
- Clustering
- Results
- Linking to explanatory variables
- Problems

# Data

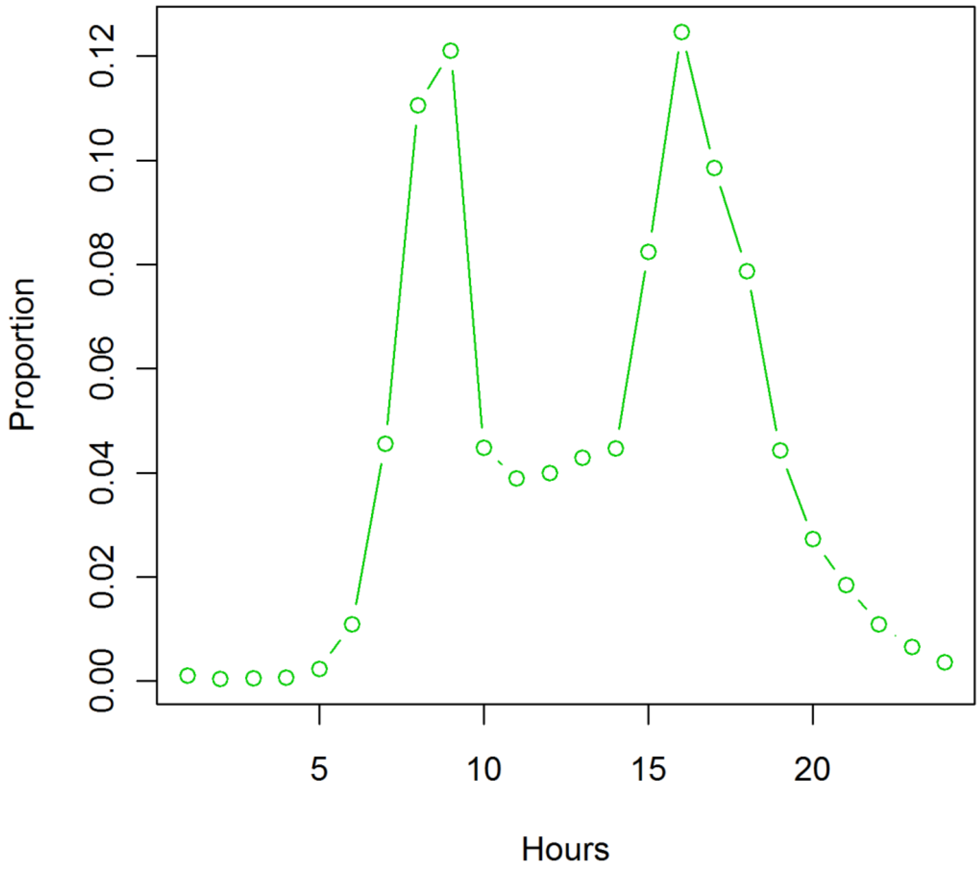
- From Sustrans, a charity that promotes sustainable transport in the UK
- Responsible for planning and delivering the National Cycle Network
- Counters count bikes!

# Counters

# Usage profiles

- What proportion of daily count per hour?

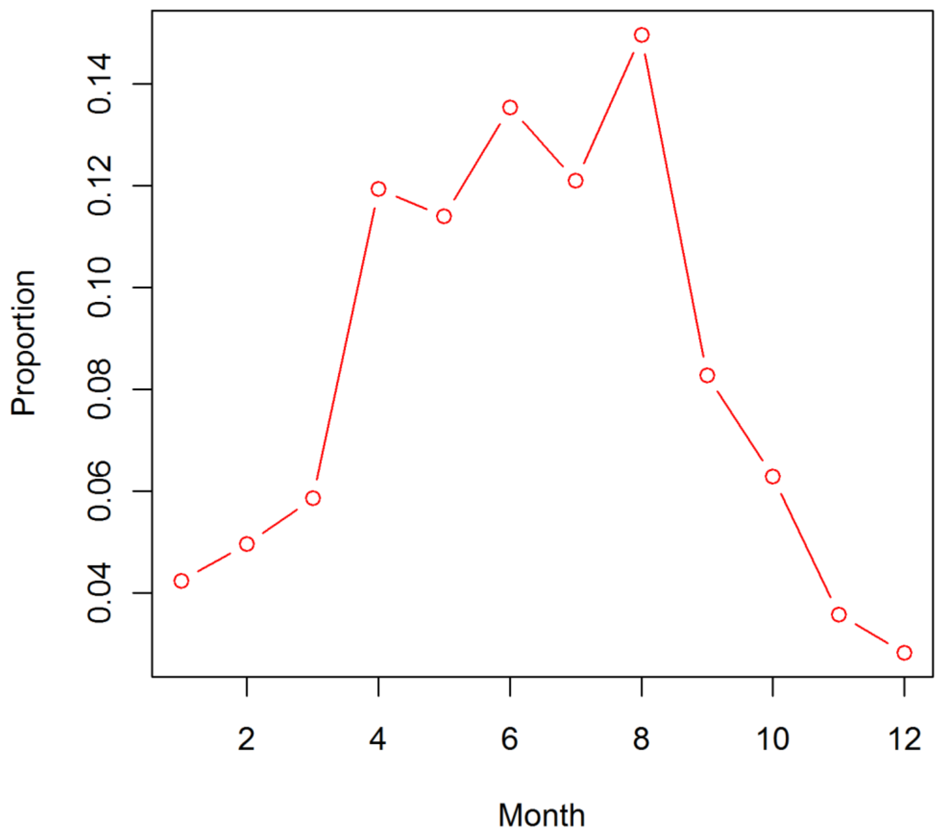
# Usage profiles



# Usage profiles

- What proportion of daily count per hour?
- What proportion of year count per month?

# Usage profiles



# Usage profiles

- What proportion of daily count per hour?
- What proportion of year count per month?
- What shape do these profiles take?

# Clustering

- Try to find common shapes.

- How do we assess dissimilarity?
  - Euclidean

    $$d(x,y)=\left(\sum_{i=1}^{n}(x_i-y_i)^2\right)^{1/2}$$

    Manhattan

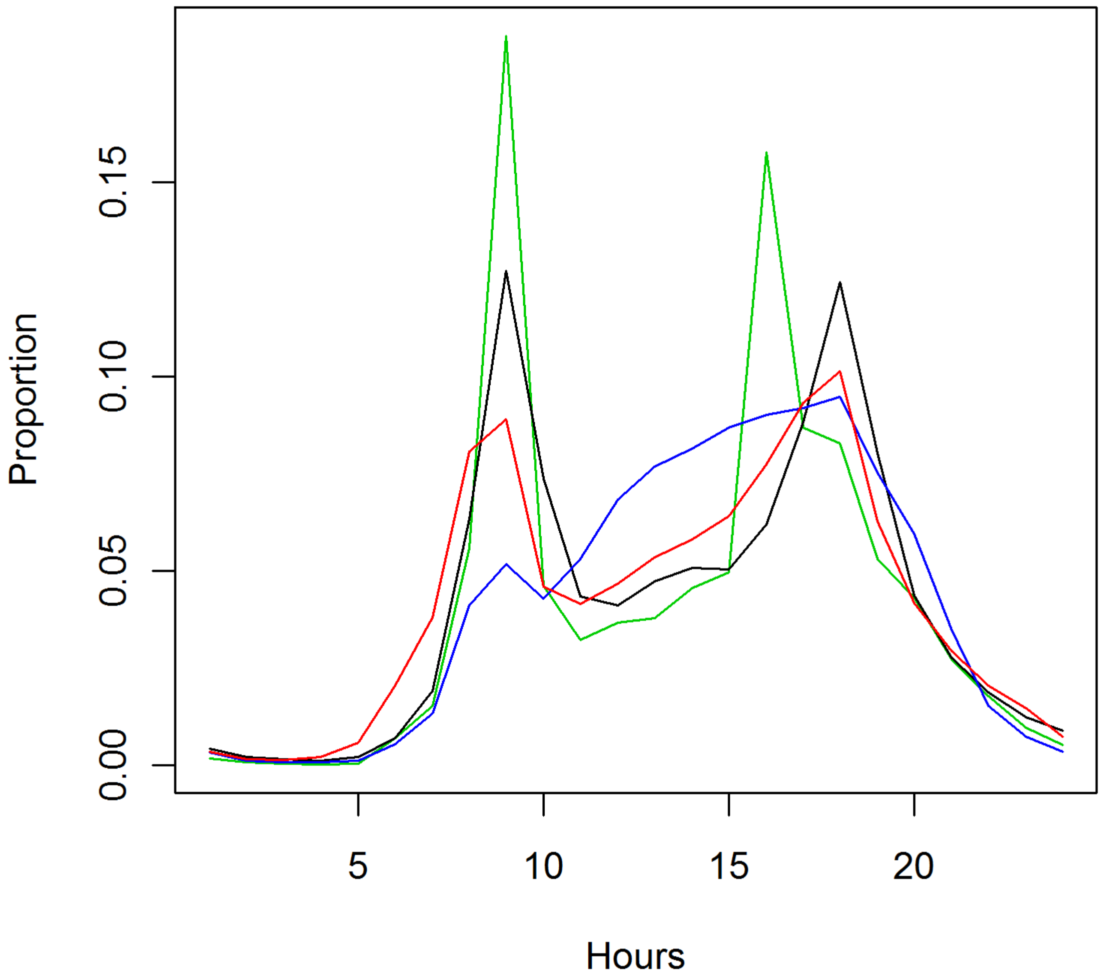
    $$d(x,y)=\sum_{i=1}^{n}|x_i-y_i|$$

    Minkowski

    $$d(x,y)=\left(\sum_{i=1}^{n}(x_i-y_i)^p\right)^{1/p}$$

- K-means clustering on daily profiles.

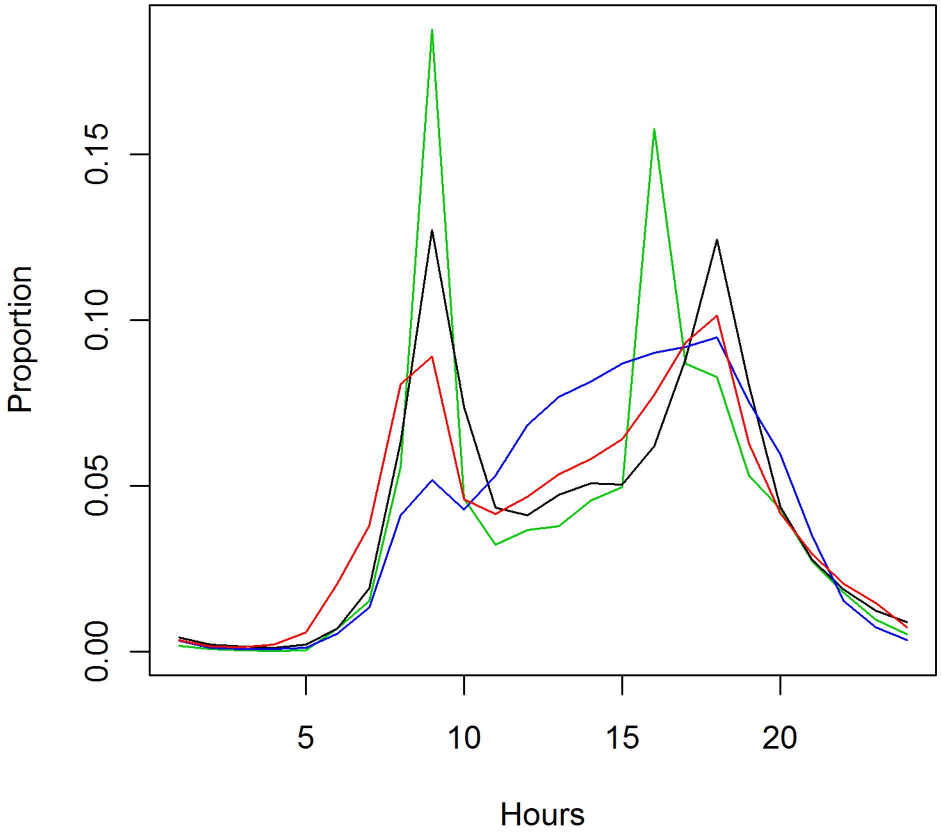
# Result



# The Statistics of Cycling

## Result!

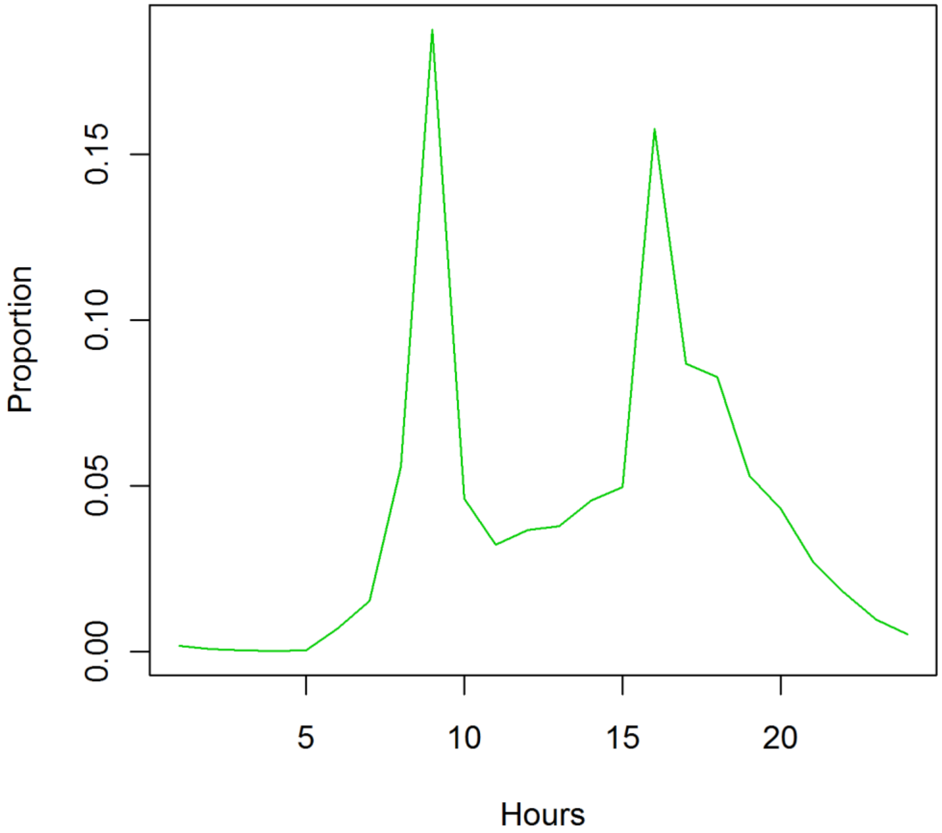
- 4 shapes!



# The Statistics of Cycling

## Result!

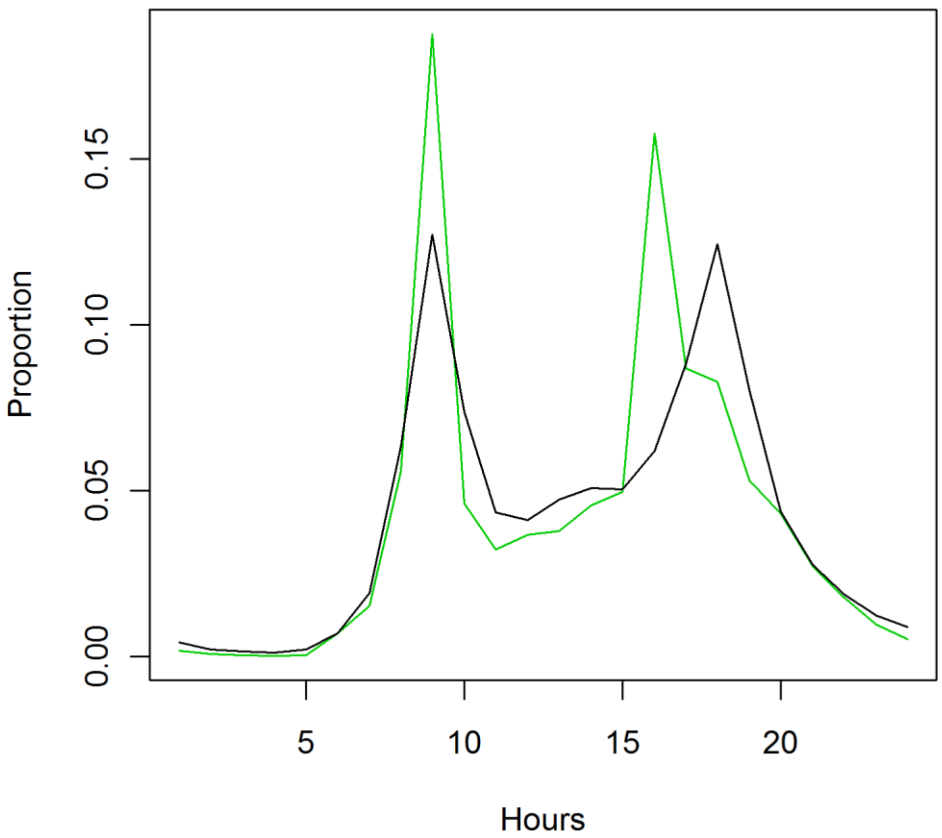
- 4 shapes!
- Schools



# Result!

- 4 shapes!

- Schools
- Commuter

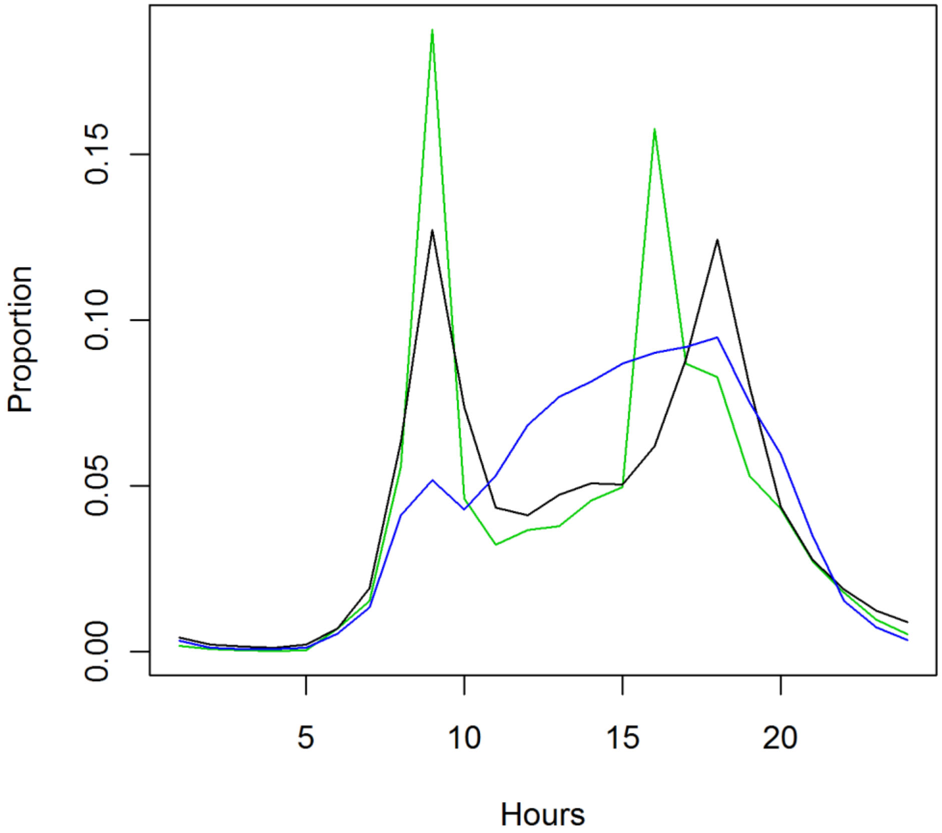


# The Statistics of Cycling

## Result!

- 4 shapes!

- Schools
- Commuter
- Leisure

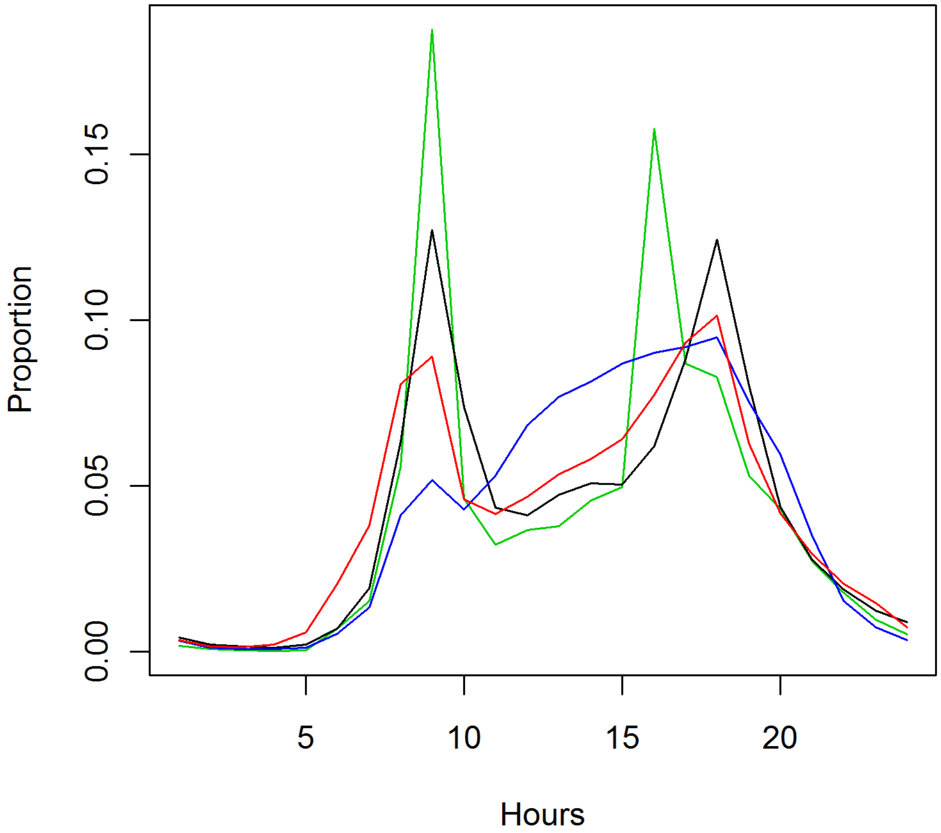


# The Statistics of Cycling

## Result!

- 4 shapes!

- Schools
- Commuter
- Leisure
- Hybrid (shopping?)



# Relate to explanatory variables

- Responses to Sustrans counter location information
- Try to fit a multinomial logit model (MLM) to "predict" classification.
  - A multivariate generalised linear model

$$\boldsymbol{\mu}_i = E[\mathbf{Y}_i] = (\pi_1(\mathbf{x}_i), \ldots, \pi_{J-1}(\mathbf{x}_i))$$

$$\mathbf{g}(\boldsymbol{\mu}_i) = \alpha_i + \mathbf{X}_i \boldsymbol{\beta}$$

$$g_j(\boldsymbol{\mu}_i) = \log \frac{\mu_{ij}}{1 - (\mu_{i1} + \ldots + \mu_{i,J-1})}$$

# Relate to explanatory variables

- Specifically baseline category logit models
- Choose one category as a baseline
  - Modal category, or just the first/last one
- We compare other categories to the baseline
- Fit $\boldsymbol{\beta}$ using maximum likelihood estimation

# Relate to explanatory variables

- Response probabilities

$$\pi_j(\mathbf{x}) = \frac{\exp(\alpha_j + \boldsymbol{\beta}_j^T \mathbf{x})}{1 + \sum_{k=1}^{J-1} \exp(\alpha_k + \boldsymbol{\beta}_k^T \mathbf{x})}$$

$$\alpha_J = 0$$

$$\boldsymbol{\beta}_J = 0$$

# Relate to explanatory variables

- Fit the following model

```
classification ~ Trafficfreeroute + region
```

- Table of observed responses

| region | midlands | | north | | south | |
|---|---|---|---|---|---|---|
| Trafficfreeroute | 0 | 1 | 0 | 1 | 0 | 1 |
| classification | | | | | | |
| commuter | 1 | 0 | 0 | 8 | 13 | 5 |
| hybrid | 12 | 12 | 2 | 14 | 2 | 1 |
| leisure | 0 | 7 | 2 | 16 | 1 | 6 |
| schools | 1 | 0 | 0 | 2 | 0 | 2 |

# Relate to explanatory variables

- Response probabilities, say we wanted to know how we might classify a counter in the North that is traffic free.

```
  commuter      hybrid    leisure    schools
0.16622310 0.34772490 0.43838972 0.04766228
```

# Problems

- If there is a zero in the table of observed responses, then parameter estimation sometimes breaks down.

- Limited data

- Schools result is not explained by any of the explanatory variables

# Questions?

If you worried about falling off the bike,
you’d never get on.

Lance Armstrong

# Example of parameter estimation failing

- Route adjacent to road table

| | commuter | hybrid | leisure | schools |
|---|---|---|---|---|
| 0 | 17 | 28 | 32 | 4 |
| 1 | 10 | 15 | 0 | 1 |

- Traffic free route table

| | commuter | hybrid | leisure | schools |
|---|---|---|---|---|
| 0 | 14 | 16 | 3 | 1 |
| 1 | 13 | 27 | 29 | 4 |

# Example of parameter estimation failing

```
multinom(formula = classification ~
route, data = newClassRoute)

Coefficients:
         (Intercept)        route
hybrid     0.4989866  -0.09345878
leisure    0.6324810 -10.71041396
schools   -1.4469457  -0.85564786

Std. Errors:
         (Intercept)      route
hybrid     0.3074673  0.5110844
leisure    0.3001218 48.8004645
schools    0.5557196  1.1869667
```

```
multinom(formula = classification ~
Trafficfreeroute, data = newClassAll)

Coefficients:
         (Intercept) Trafficfreeroute
hybrid     0.1335310        0.5973803
leisure   -1.5404378        2.3427939
schools   -2.6390253        1.4603639

Std. Errors:
         (Intercept) Trafficfreeroute
hybrid     0.3659628        0.4978849
leisure    0.6362076        0.7184476
schools    1.0350836        1.1825086
```