The Magic of Multiple Linear Regression

Joe Guinness

August 19, 2025

Department of Statistics & Data Science

General Presentation Tips

Avoid big blocks of text. Avoid long sentences.

Proofread, especially titles.

Use figures or pictures. Plan what to say while on screen.

Take time to explain equations. Difficult for even expert audiences to decipher.

Practice. Film yourself (not as scary as it sounds).

Speak loudly and clearly.

If you're comfortable, audience is comfortable.

What is a Statistical Analysis?

The following workflow is useful for organizing my thoughts:

1. Formulate a question.
2. Collect data relevant to the question.
3. Specify a statistical model for the data.
4. Use the data to estimate model parameters.
5. Make a judgment about the answer to the question.

Nike Vaporfly Running Shoes

1. Formulate a Question

This is often harder than it sounds.

Very important skill to develop for statisticians and data scientists

1. Formulate a Question

This is often harder than it sounds.

Very important skill to develop for statisticians and data scientists

Do vaporfly shoes make you run faster?

1. Formulate a Question

This is often harder than it sounds.

Very important skill to develop for statisticians and data scientists

Do vaporfly shoes make you run faster?

If you run the same race twice, would you run faster in the vaporflys?

1. Formulate a Question

This is often harder than it sounds.

Very important skill to develop for statisticians and data scientists

Do vaporfly shoes make you run faster?

If you run the same race twice, would you run faster in the vaporflys?

If a literal clone of you ran in vaporflys on the same day, would the clone run faster?

2. Collect Data Relevant to the Question

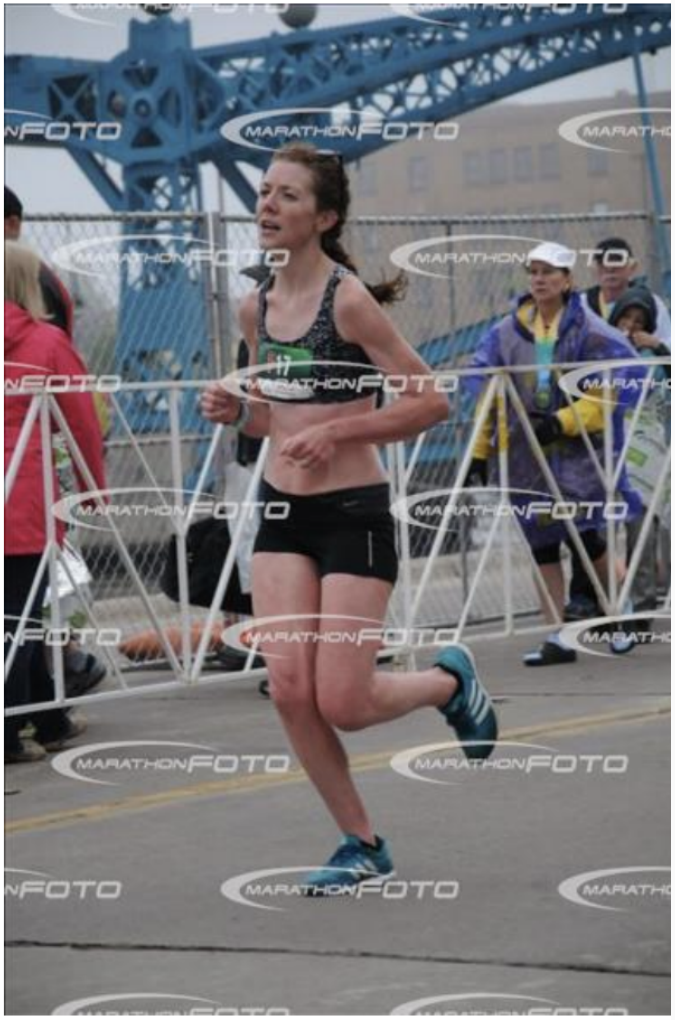
Need a dataset with marathon performances
(names, races, dates, finish time, etc.)

- Scraped from marathonguide.com

Need a dataset of what shoes runners wore in these races

- Doesn't exist.
- Need to do some grunt work to get this information

2. Collect Data Relevant to the Question



2. Collect Data Relevant to the Question

Goal is to see the effect of vaporflys on elite athletes in particular

So we selected runners who met performance standards:

1. Men: 2:24 or better
2. Women: 2:45 or better

However, there is a subtle pitfall here.

Avoid overselecting for athletes that benefit from vaporflys most

2. Collect Data Relevant to the Question

Goal is to see the effect of vaporflys on elite athletes in particular

So we selected runners who met performance standards:

1. Men: 2:24 or better in 2015 or 2016
2. Women: 2:45 or better in 2015 or 2016

However, there is a subtle pitfall here.

Avoid overselecting for athletes that benefit from vaporflys most

2. Collect Data Relevant to the Question

Data Table:

i	name	$j(i)$	race	$k(i)$	s_i	v_i	y_i
1	Beth	1	Chicago 2018	1	female	0	2:42:30
2	Beth	1	Boston 2019	2	female	1	2:37:25
3	Jim	2	Chicago 2018	1	male	0	2:18:11
4	Jim	2	Boston 2019	2	male	0	2:20:45
5	Dave	3	New York 2016	3	male	0	2:20:45
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

3. Specify a Statistical Model for the Data

What is a statistical model?

3. Specify a Statistical Model for the Data

What is a statistical model?

Set of mathematical equations meant to encode the random and non-random processes that produced the data.

3. Specify a Statistical Model for the Data

What is a statistical model?

Set of mathematical equations meant to encode the random and non-random processes that produced the data.

Allows us to make inferences that come with uncertainties

3. Specify a Statistical Model for the Data

What is a statistical model?

Set of mathematical equations meant to encode the random and non-random processes that produced the data.

Allows us to make inferences that come with uncertainties

If runners who wore the vaporflys ran 3 minutes faster on average than those who didn't, it makes a big difference whether there were 4 runners versus 400 runners in the sample.

3. Specify a Statistical Model for the Data

What is a statistical model?

Set of mathematical equations meant to encode the random and non-random processes that produced the data.

Allows us to make inferences that come with uncertainties

If runners who wore the vaporflys ran 3 minutes faster on average than those who didn't, it makes a big difference whether there were 4 runners versus 400 runners in the sample.

More subtly, if the set of runners who wore vaporflys is totally separate from the set of runners who didn't, selection bias could explain the results.

3. Specify a Statistical Model for the Data

$$Y_i = b_0 + b_1 v_i + A_{j(i)} + B_{k(i)} + \varepsilon_i$$

3. Specify a Statistical Model for the Data

$$Y_i = b_0 + b_1 v_i + A_{j(i)} + B_{k(i)} + \varepsilon_i$$

i = label for the performance

$v_i = 0$ (no vaporfly in performance i)

$v_i = 1$ (vaporfly in performance i)

$j(i)$ = label for the runner in performance i

$k(i)$ = label for the race in performance i

$A_j =$ ability of runner j

$B_k =$ difficulty of race k

$\varepsilon_i =$ random error

3. Specify a Statistical Model for the Data

Hypothetical performances with and without the vaporfly

$$Y_1 = b_0 + b_1 * 0 + A_1 + B_1 + \varepsilon_1 \text{ (no vaporfly)}$$
$$Y_2 = b_0 + b_1 * 1 + A_1 + B_1 + \varepsilon_2 \text{ (vaporfly)}$$

$$Y_2 - Y_1 = b_1 + (\varepsilon_2 - \varepsilon_1)$$

b_1 tells us how much we expect the performances to differ

4. Use the data to estimate model parameters

Data Table:

i	name	$j(i)$	race	$k(i)$	s_i	v_i	y_i
1	Beth	1	Chicago 2018	1	female	0	2:42:30
2	Beth	1	Boston 2019	2	female	1	2:37:25
3	Jim	2	Chicago 2018	1	male	0	2:18:11
4	Jim	2	Boston 2019	2	male	0	2:20:45
5	Dave	3	New York 2016	3	male	0	2:20:45
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

4. Use the data to estimate model parameters

$$Y_1 = b_0 + b_1 * 0 + A_1 + B_1 + \varepsilon_1$$
$$Y_2 = b_0 + b_1 * 1 + A_1 + B_2 + \varepsilon_2$$
$$Y_3 = b_0 + b_1 * 0 + A_2 + B_1 + \varepsilon_3$$
$$Y_4 = b_0 + b_1 * 0 + A_2 + B_2 + \varepsilon_4$$

4. Use the data to estimate model parameters

$$Y_1 = b_0 + b_1 * 0 + A_1 + B_1 + \varepsilon_1$$
$$Y_2 = b_0 + b_1 * 1 + A_1 + B_2 + \varepsilon_2$$
$$Y_3 = b_0 + b_1 * 0 + A_2 + B_1 + \varepsilon_3$$
$$Y_4 = b_0 + b_1 * 0 + A_2 + B_2 + \varepsilon_4$$

How about this?

$$Y_2 - Y_1 = b_1 + (B_2 - B_1) + (\varepsilon_2 - \varepsilon_1)$$

4. Use the data to estimate model parameters

$$Y_1 = b_0 + b_1 * 0 + A_1 + B_1 + \varepsilon_1$$
$$Y_2 = b_0 + b_1 * 1 + A_1 + B_2 + \varepsilon_2$$
$$Y_3 = b_0 + b_1 * 0 + A_2 + B_1 + \varepsilon_3$$
$$Y_4 = b_0 + b_1 * 0 + A_2 + B_2 + \varepsilon_4$$

How about this?

$$Y_2 - Y_1 = b_1 + (B_2 - B_1) + (\varepsilon_2 - \varepsilon_1)$$

This is better

$$(Y_2 - Y_1) - (Y_4 - Y_3) = b_1 + \varepsilon_2 - \varepsilon_1 + \varepsilon_3 - \varepsilon_4$$

4. Use the data to estimate model parameters

How to construct estimates from hundreds of performances?

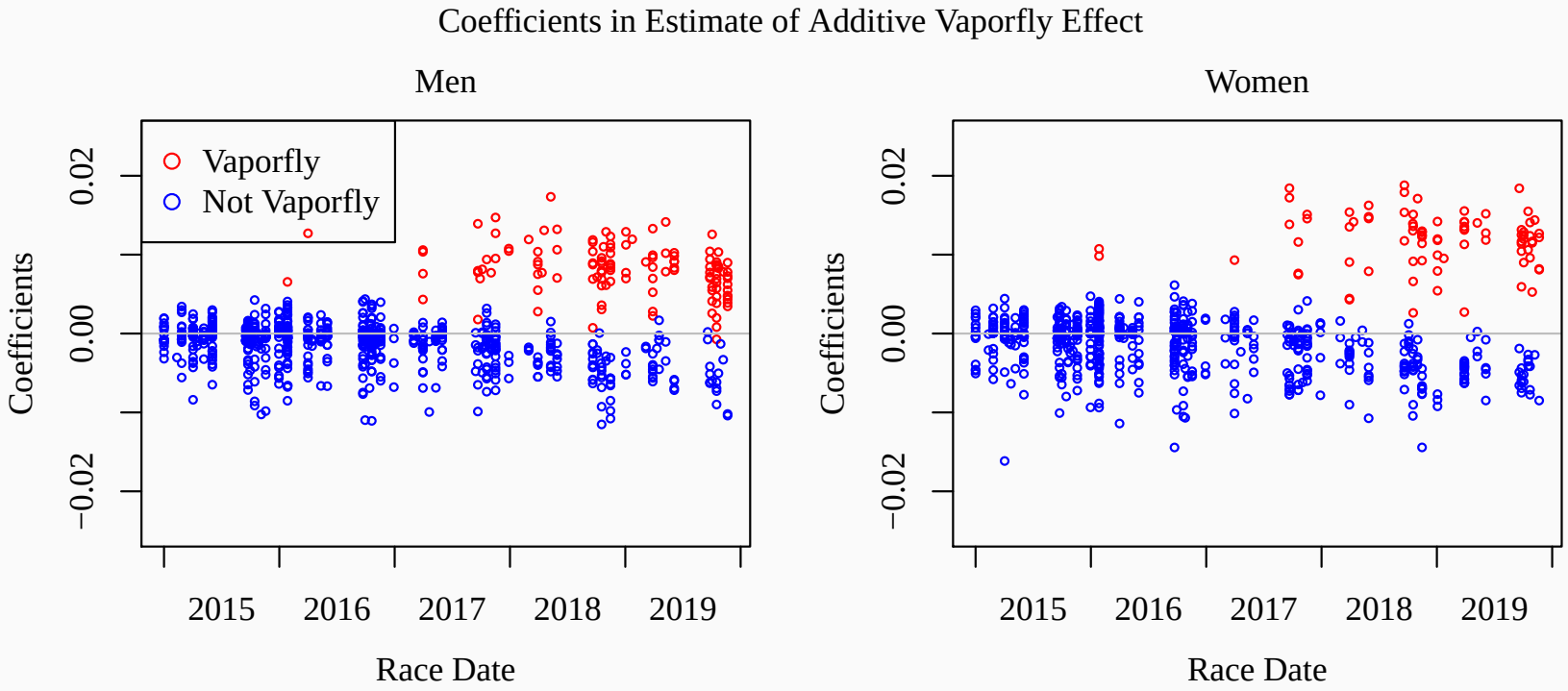
This is the magic of multiple linear regression.

It figures out the best combination of the observations for estimating the parameters in the model.

Exact formula uses matrix inverses and multiplication

You would learn this in SDS 4130 - *Linear Statistical Models*

4. Use the data to estimate model parameters



4. Use the data to estimate model parameters

Here are the results:

	men minutes	women minutes
	estimate (s.e.)	estimate (s.e.)
b_0	139.69 (0.59)	159.83 (0.81)
b_1	**−2.95 (0.60)**	**−2.18 (0.81)**
σ_1	4.175	6.40
σ_2	1.852	2.33
σ_3	1.874	2.43
σ_4	4.108	5.02

5. Make a judgment about the answer to the question

Depends on several factors:

Has the data collection introduced any biases?

Are the model assumptions appropriate?

The raw statistical results - is the result significant?

Is the result plausible given other evidence?

Statistics and Data Science at Wash U

Statistics Major:

Emphasizes probability models, statistical techniques, programming, linear models, critical thinking, statistical intuition

Data Science Major:

Broader, includes statistics, more programming, database management, computer science

My contact: joeguinness@wustl.edu

Reference

An Observational Study of the Effect of Nike Vaporfly Shoes on Marathon Performance

Joseph Guinness, Debasmita Bhattacharya, Jenny Chen, Max Chen, Angela Loh

https://arxiv.org/abs/2002.06105v2