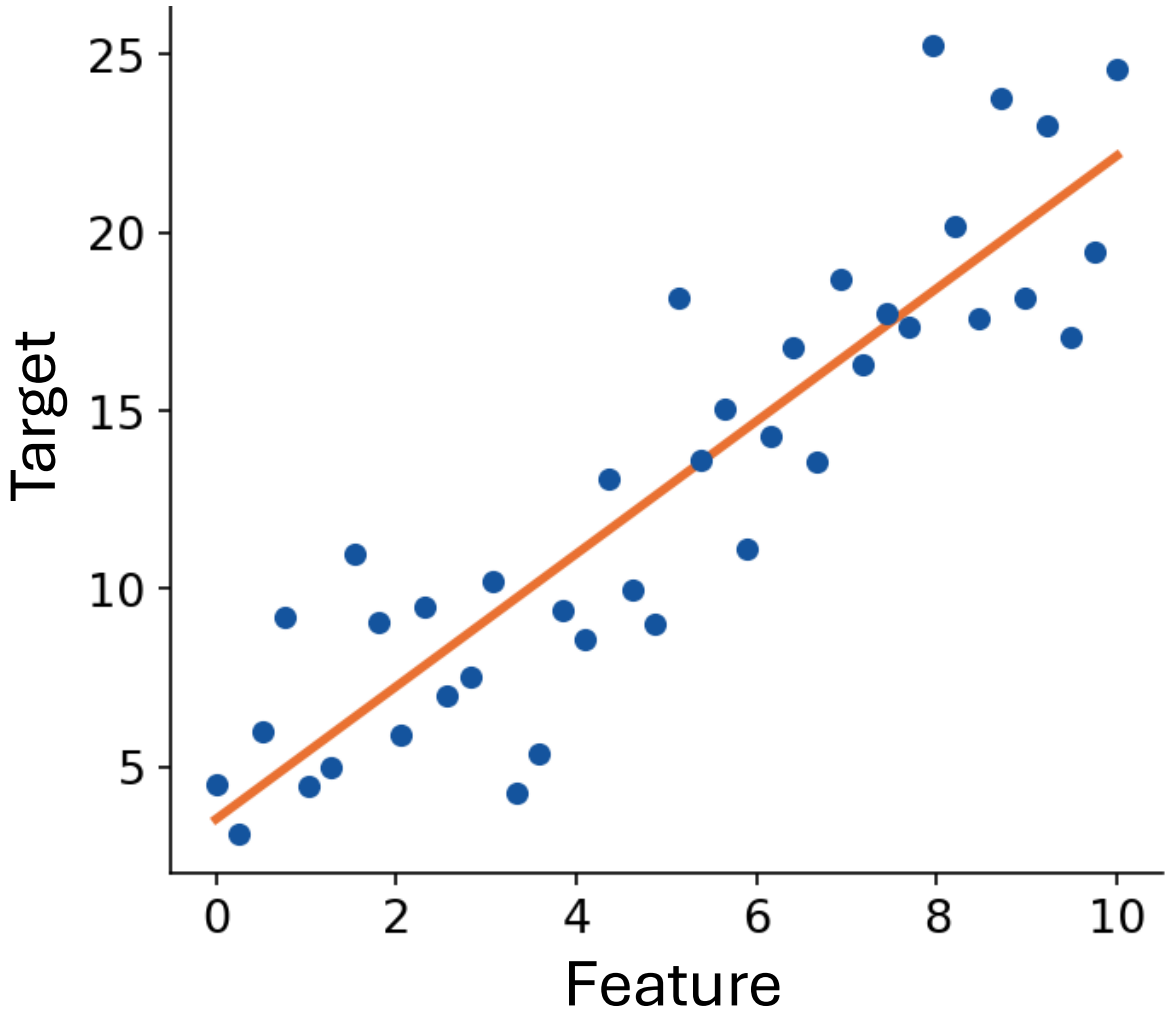
COM SCI M148, Lecture 4:
Multiple Linear & Polynomial Regression

Prof. Nathan Szymanski
nszymanski@seas.ucla.edu
Department of Materials Science

UCLA

Samueli
School of Engineering

Recall: Linear regression



$$\hat{y} = \beta_0 + \beta_1 x$$

- We built a linear model to predict a **target ($\hat{y}$)** from a **single variable (x)**
- This required the fitting of just **2 variables (β_0 and β_1)** by minimizing MSE loss

The real world often involves *many* variables

Your exam grade depends on more than just study time (sleep, lecture attendance, *etc.*)







Sales of a product depends on advertising across > 1 medium


Radio


TV

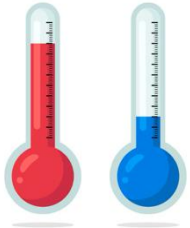


Newspaper

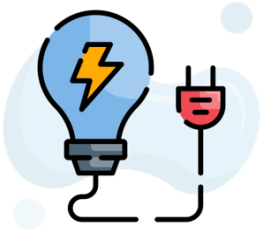
The properties of cement depend on its water and sand content, curing temperature, *etc.*







The energy use of a building depends on building size, outside temperature, number of occupants, *etc.*







How to deal with multiple variables

Sales of a product depends on advertising through different venues



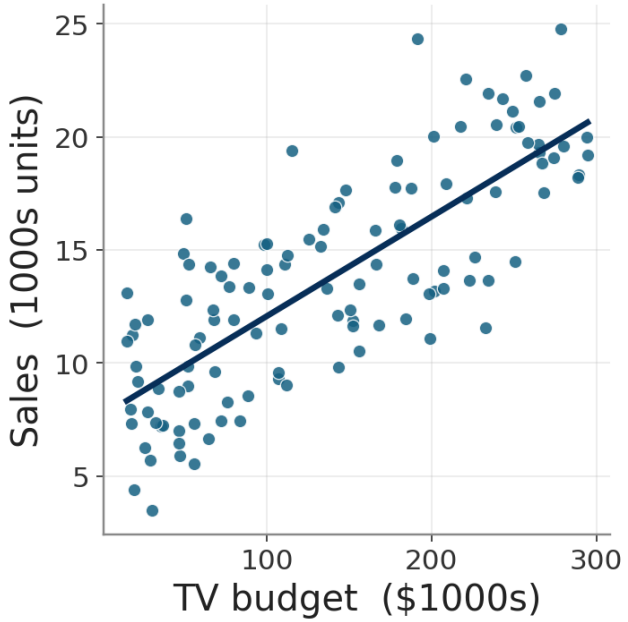


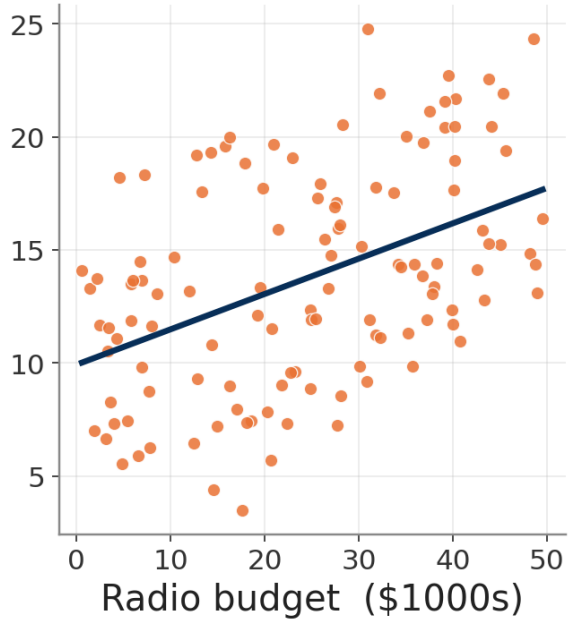


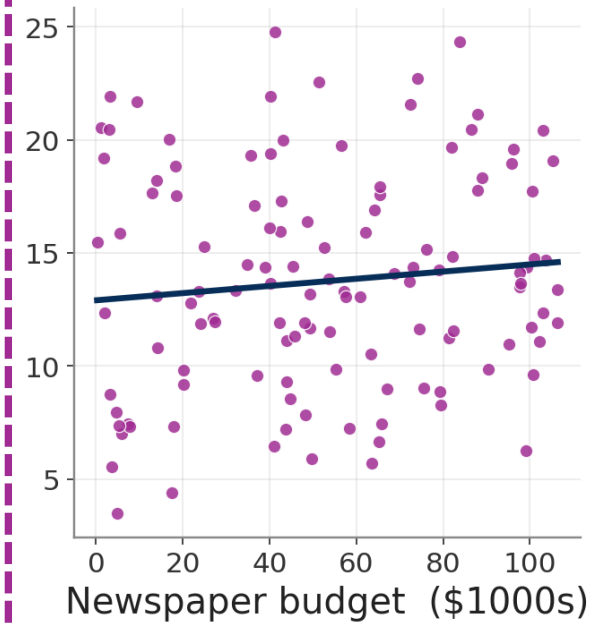
TV budget	Radio budget	Paper budget	Unit sales
$100k	$8k	$59k	11k
$80k	$25k	$18k	9k
$195k	$42k	$92k	18k
$232k	$21k	$5k	23k

We could model each variable **one at a time**, but this would not account for the effects of the other variables.

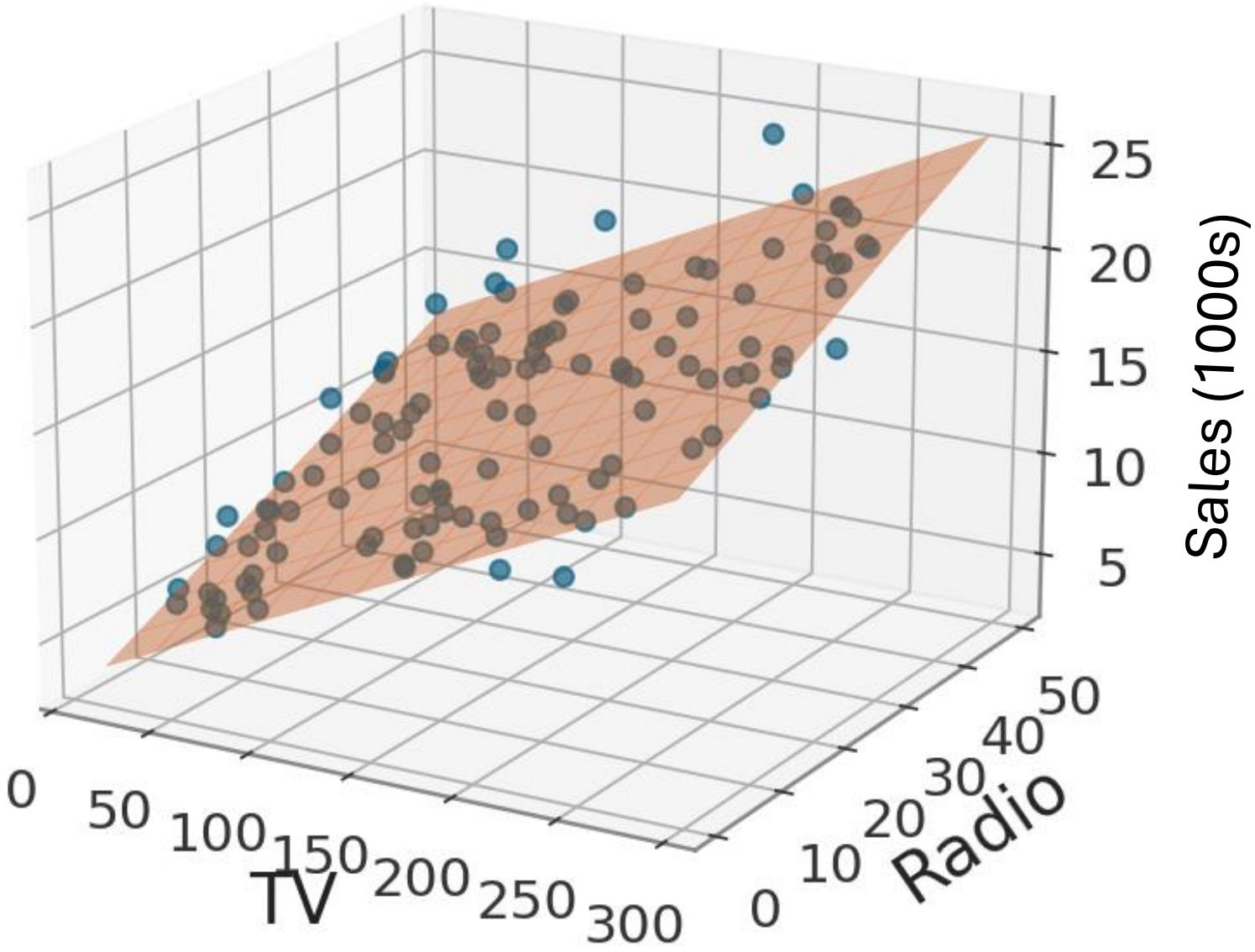
Large unexplained variation







The variability is explained by other features



How do we find the plane that best fits our data?

Multiple Linear Regression

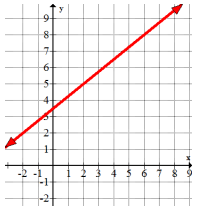
*Predicting one target from **multiple input features***

Adding dimensions to our linear model

1 variable (x)

$$\hat{y} = \beta_0 + \beta_1 x$$

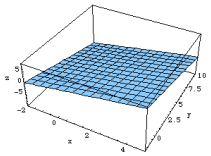
A line



2 variables (x_1 and x_2)

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

A plane



n variables ($x_1, x_2, \ldots, x_n$)

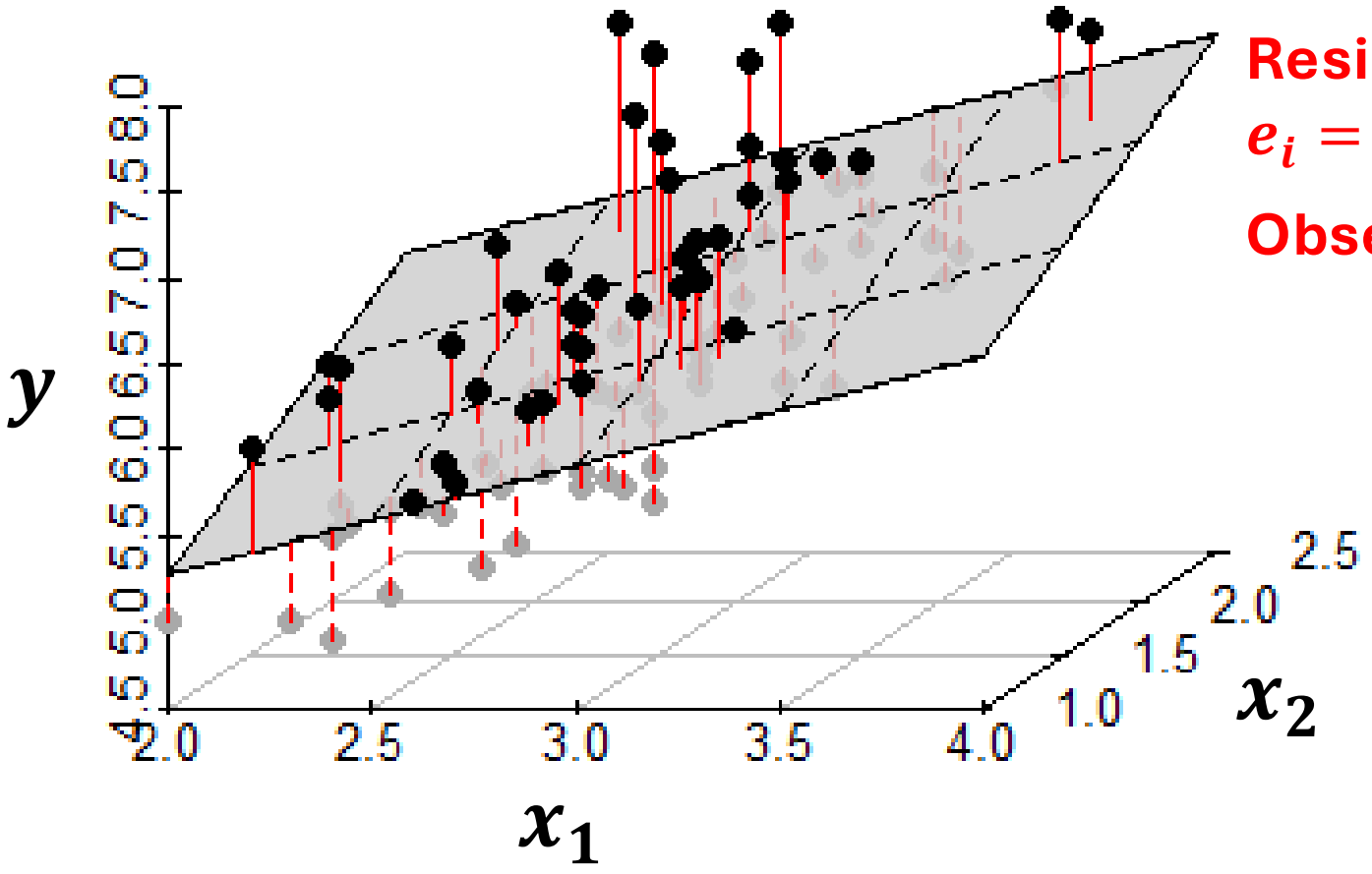
$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

?

- Each case is still a **linear problem:** the target is just a weighted sum of independent variables
- With n **variables, we have** $n + 1$ **coefficients** that need optimized: one *y*-intercept (β_0) + one slope for each predictor ($\beta_{i\neq 0}$)
- Luckily, the same loss and fitting "machinery" we developed last time still apply here!

Choose β_i that minimize the MSE loss

Residuals in higher dimensions



Residual:

$e_i = y_i - \hat{y}_i$

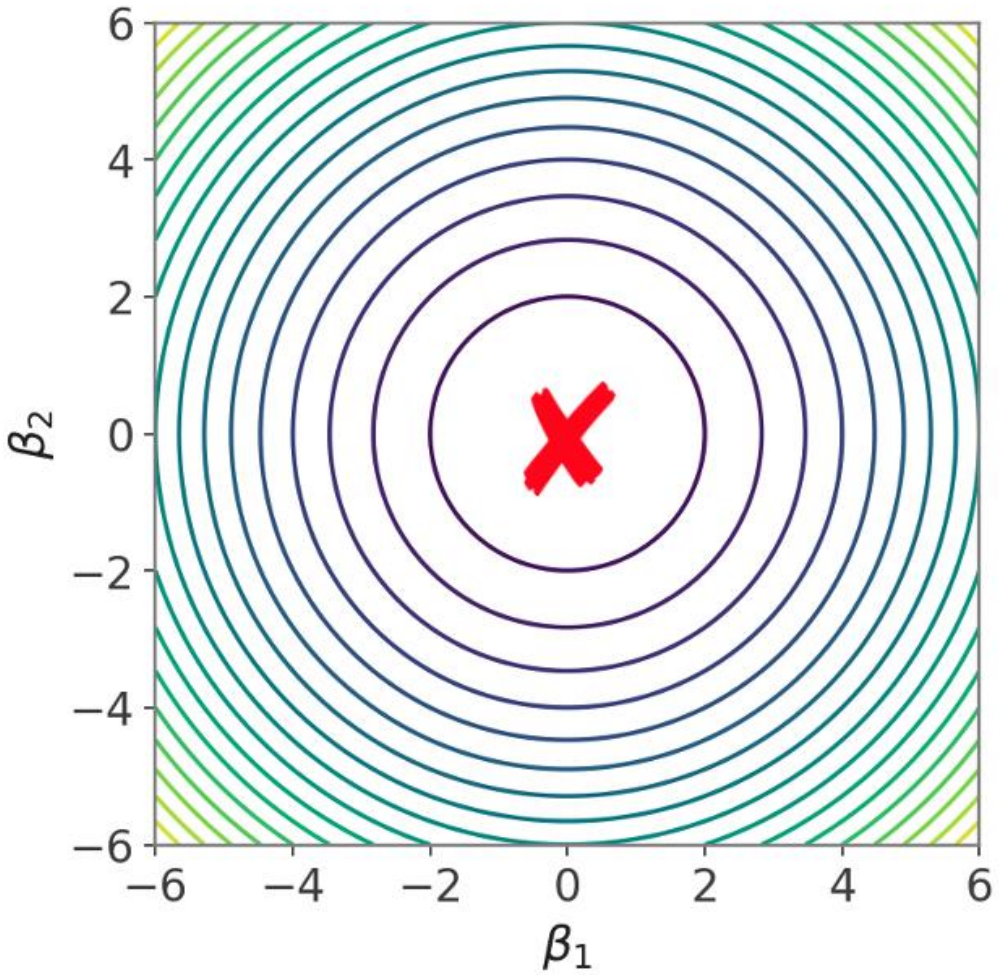
Observed – Predicted

The Mean Squared Error (MSE):

$$\mathcal{L} = \frac{1}{n}\sum_{i=1}^{n} |e_i|^2$$

Find the values of β_i where $\nabla\mathcal{L} = 0 \rightarrow$ the loss is minimal

Residuals in higher dimensions



$$\nabla\mathcal{L} = \begin{pmatrix} \frac{\partial\mathcal{L}}{\partial\beta_0} \\ \frac{\partial\mathcal{L}}{\partial\beta_1} \\ \cdots \\ \frac{\partial\mathcal{L}}{\partial\beta_n} \end{pmatrix} = \mathbf{0}$$

The gradient

Set all derivates equal to zero and solve

We've already done this math! It's the same as last time. Now in matrix form…

Find the values of β_i where $\nabla\mathcal{L} = 0 \rightarrow$ the loss is minimal

Normal equation for least-squares regression

In the previous lecture, we derived:

$$\beta_0 = \bar{y} - \beta_1 \bar{x} \qquad \beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

By solving:

$$\frac{d\mathcal{L}}{d\beta_0} = 0, \qquad \frac{d\mathcal{L}}{d\beta_1} = 0$$

And then we (hand-wavily) said this generalizes to matrix form:

$$\hat{\beta} = \left(X^{\mathrm{T}}X\right)^{-1}X^{\mathrm{T}}y$$

Where X **is a matrix of our predictors** and y **is the response vector**

→ ***Let's prove it!***

Normal equation derivation

If you have p predictors and n observations, then $\boldsymbol{X}$ becomes an $n \times (p+1)$ matrix.

n rows
$p+1$ columns

n outputs (predicted)

$p+1$ coefficients

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \quad \hat{\boldsymbol{y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \dots \\ \hat{y}_n \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_p \end{bmatrix}$$

Each row is one sample/observation

$$X\beta = \begin{pmatrix} \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p} \\ \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2p} \\ \dots \\ \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_p x_{np} \end{pmatrix} = \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \dots \\ \hat{y}_n \end{pmatrix} = \hat{y}$$

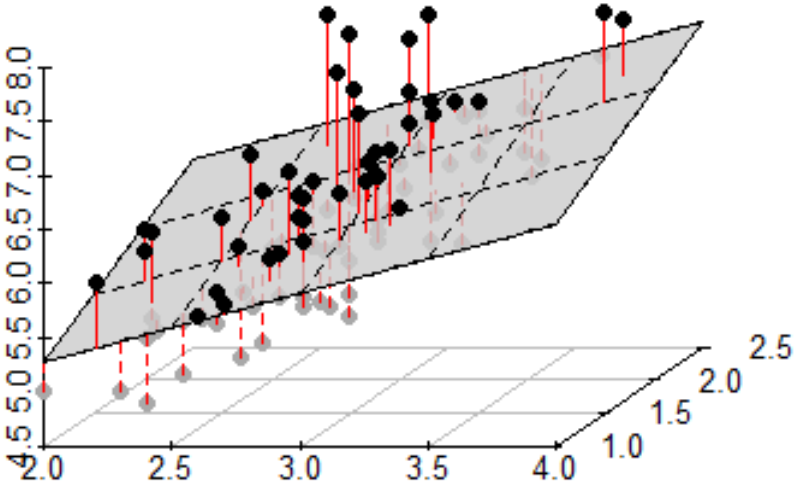
Each column is one feature

This is just a system of linear equations written in matrix form

Normal equation derivation

Residuals: $\boldsymbol{r} = y - \hat{y} = y - X\beta$

MSE Loss: $\mathcal{L} = \|r\|^2 = \|y - X\beta\|^2$



$$\boldsymbol{r} = \begin{pmatrix} y_1 - \hat{y}_1 \\ y_2 - \hat{y}_2 \\ \dots \\ y_n - \hat{y}_n \end{pmatrix} = \begin{pmatrix} y_1 - (\beta_0 + \beta_1 x_{11} + \cdots + \beta_p x_{1p}) \\ y_2 - (\beta_0 + \beta_1 x_{21} + \cdots + \beta_p x_{2p}) \\ \dots \\ y_n - (\beta_0 + \beta_1 x_{n1} + \cdots + \beta_p x_{np}) \end{pmatrix}$$

Each row corresponds to one residual

$$X\beta = \begin{pmatrix} \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p} \\ \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2p} \\ \dots \\ \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_p x_{np} \end{pmatrix} = \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \dots \\ \hat{y}_n \end{pmatrix} = y$$

This is just a system of linear equations written in matrix form

Normal equation derivation

Residuals: $r = y - \hat{y} = y - X\beta$

MSE Loss: $\mathcal{L} = \|r\|^2 = \|y - X\beta\|^2$

$$\|y - X\beta\|^2 = (y - X\beta)^T(y - X\beta)$$

$$= \left[y_1 - (\beta_0 + \beta_1 x_{11} + \cdots + \beta_p x_{1p})\right]^2 + \cdots + \left[y_n - (\beta_0 + \beta_1 x_{n1} + \cdots + \beta_p x_{np})\right]^2$$

$$= \sum_{i=1}^{n} \left[y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})\right]^2$$

Observed

Predicted

This is just our familiar MSE equation!

Normal equation derivation

Now, let's minimize this loss using matrix math + calculus

MSE Loss: $\mathcal{L} = \|r\|^2 = \|y - X\beta\|^2$

Rewrite: $\|y - X\beta\|^2 = (y - X\beta)^T(y - X\beta) = y^Ty - \underbrace{y^TX\beta - \beta^TX^Ty}_{-2\beta^TX^Ty} + \beta^TX^TX\beta$

(EXPAND) **(FOIL)**

$d(\beta^T) = (d\beta)^T$

Minimize the loss: $\frac{\partial \mathcal{L}}{\partial \beta} = \frac{\partial}{\partial \beta}(y^Ty - 2\beta^TX^Ty + \beta^TX^TX\beta) = -2X^Ty + 2X^TX\beta = 0$

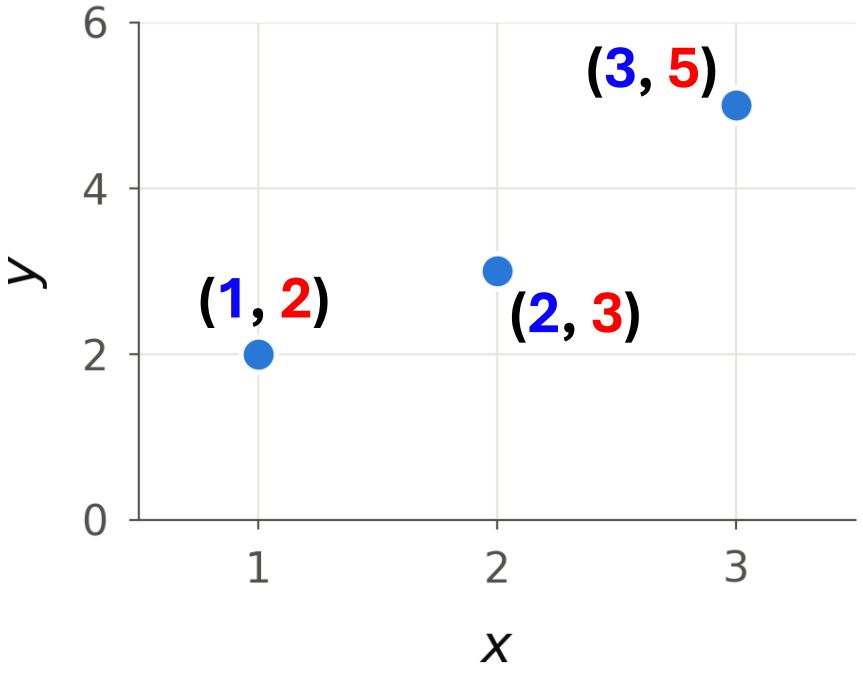
$d(\beta^TA\beta) = 2\beta A$ if A is symmetric

Solve for β by rearranging: $2X^TX\beta = 2X^Ty \rightarrow \boldsymbol{\beta = (X^TX)^{-1}X^Ty}$

Cancel the factor of 2 and multiply on the left by the inverse $(X^TX)^{-1}$

The normal equation for least-squares regression!

Using the normal equation $\beta = (X^T X)^{-1} X^T y$



$$X = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{pmatrix} \quad y = \begin{pmatrix} 2 \\ 3 \\ 5 \end{pmatrix}$$

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

This is what we want to solve for

$$X^T X = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{pmatrix} = \begin{pmatrix} 3 & 6 \\ 6 & 14 \end{pmatrix}$$

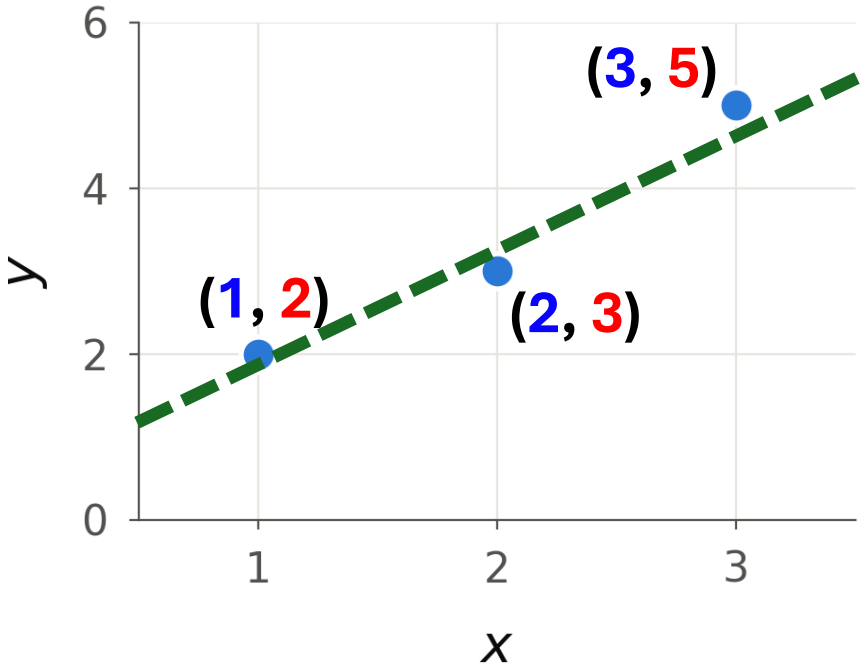
$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

Matrix inversion formula (2×2)

$$(X^T X)^{-1} = \frac{1}{3(14) - 6(6)} \begin{pmatrix} 14 & -6 \\ -6 & 3 \end{pmatrix}$$

$$\mathbf{(X^T X)^{-1} = \frac{1}{6} \begin{pmatrix} 14 & -6 \\ -6 & 3 \end{pmatrix}}$$

Using the normal equation $\beta = (X^T X)^{-1} X^T y$



$$X = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{pmatrix} \quad y = \begin{pmatrix} 2 \\ 3 \\ 5 \end{pmatrix}$$

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$ This is what we want to solve for

$$X^T y = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{pmatrix} \begin{pmatrix} 2 \\ 3 \\ 5 \end{pmatrix} = \begin{pmatrix} 10 \\ 23 \end{pmatrix}$$

$$\beta = (X^T X)^{-1} X^T y = \frac{1}{6} \begin{pmatrix} 14 & -6 \\ -6 & 3 \end{pmatrix} \begin{pmatrix} 10 \\ 23 \end{pmatrix}$$

This generalizes to many dimensions (more than one feature, like x)

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} 1/3 \\ 3/2 \end{pmatrix}$$

$$(X^T X)^{-1} = \frac{1}{6} \begin{pmatrix} 14 & -6 \\ -6 & 3 \end{pmatrix}$$

…Sorry for lots of math…

Questions?

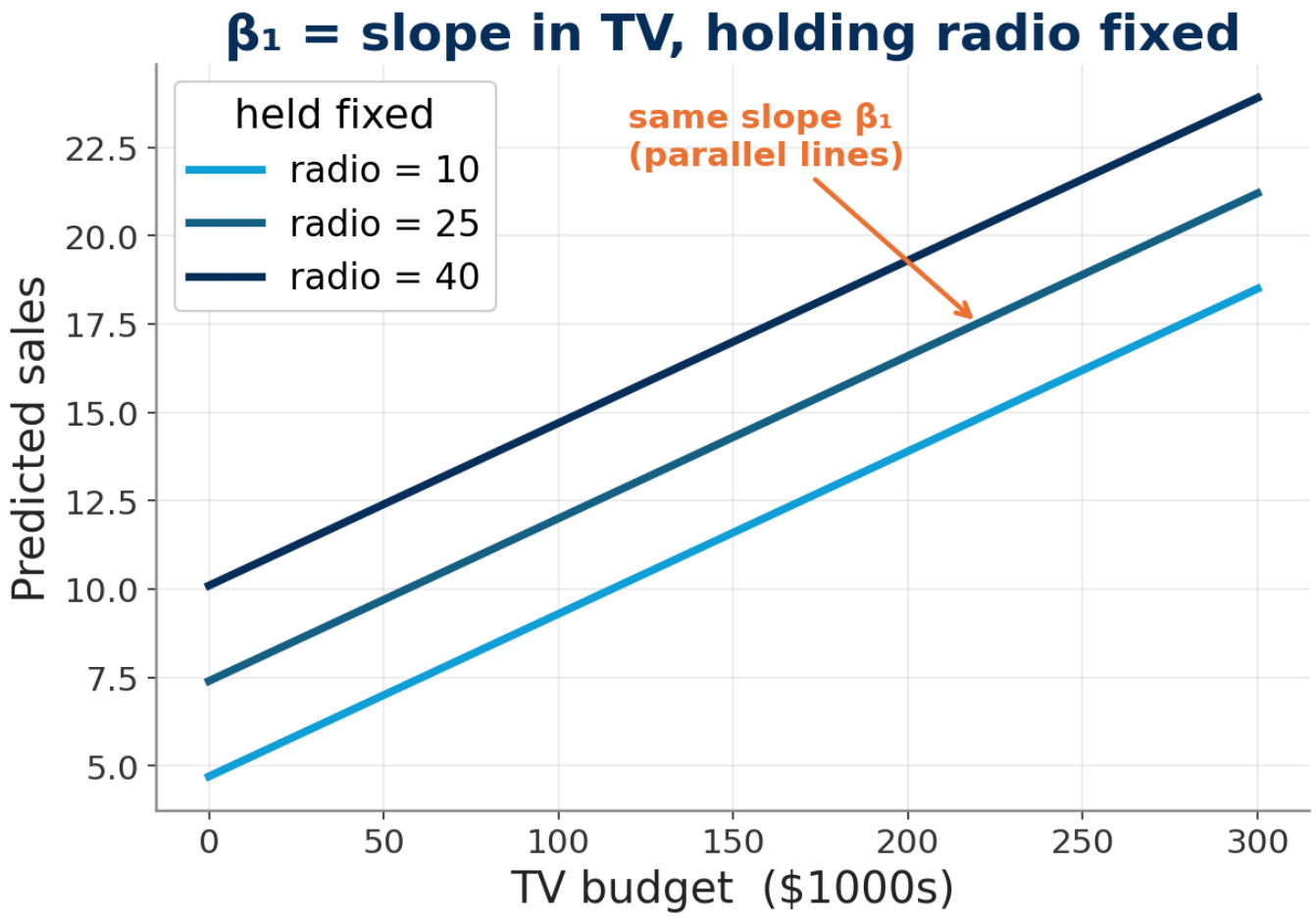
$$\boldsymbol{\beta} = \left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y}$$

Now, a break from the math: Reading the Coefficients

What does each β_i actually mean, when there are multiple features?

Each β_i represents a slope

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

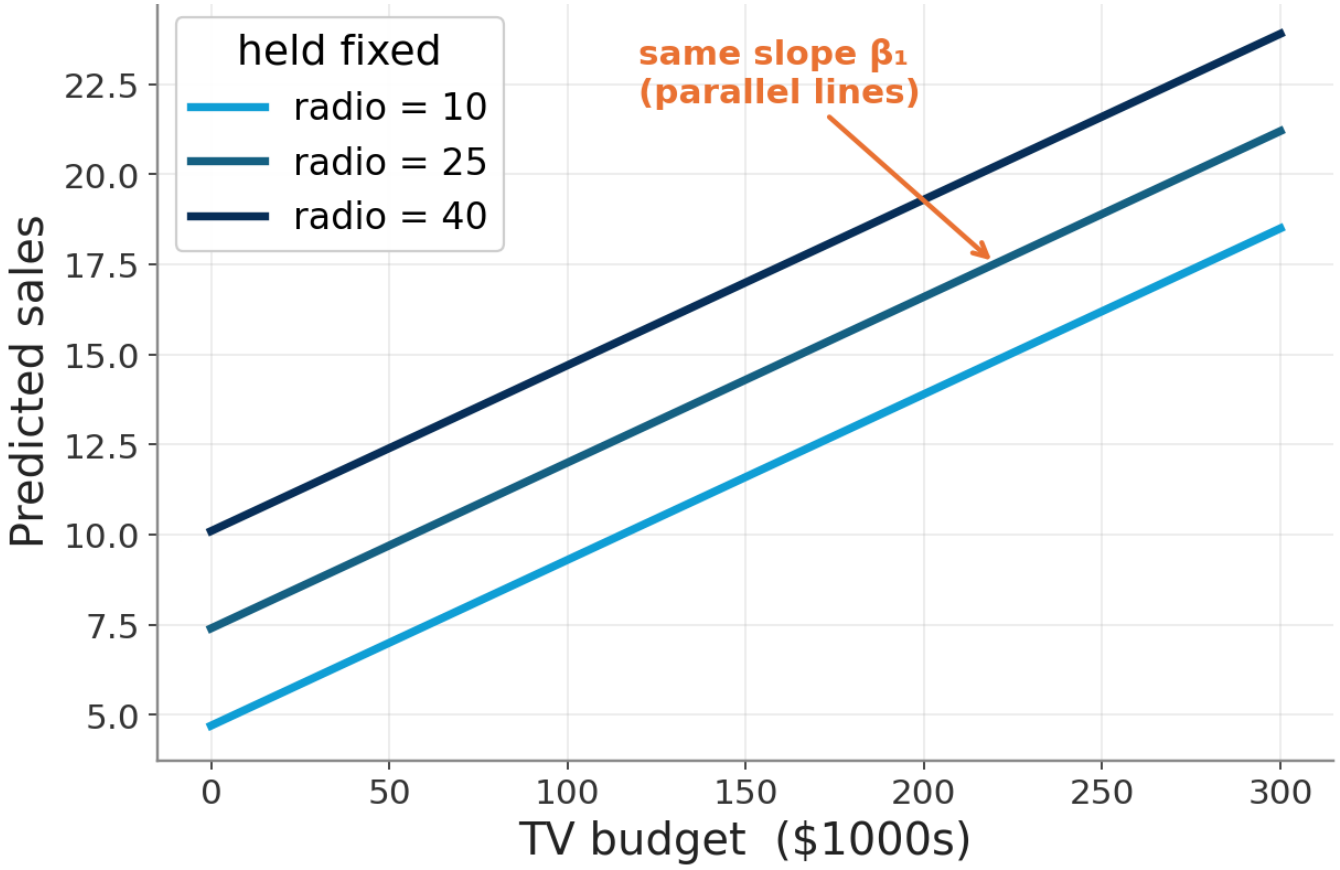


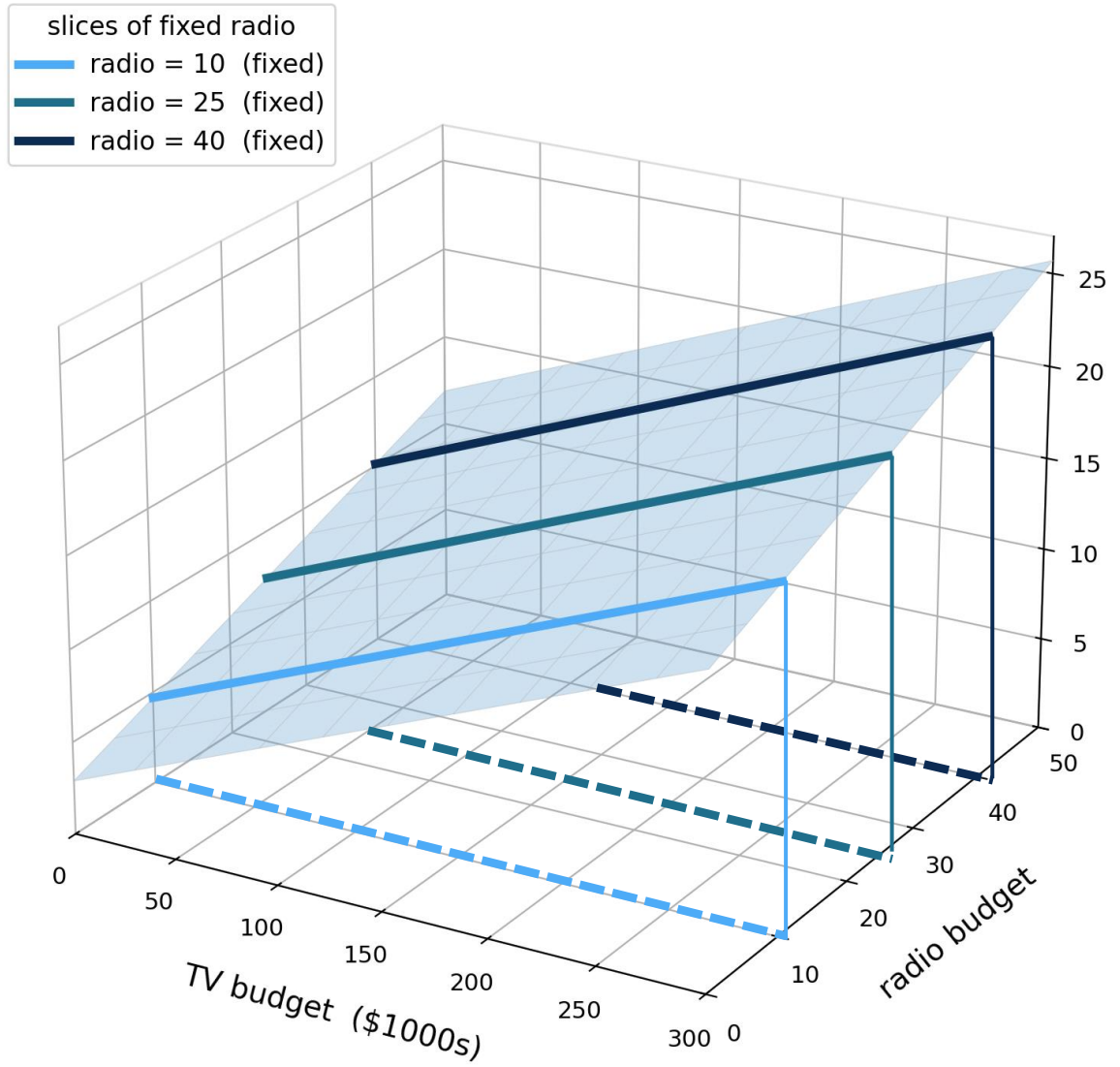
Each β_i represents the change in $\hat{y}$ per one-unit increase in x_i, **while holding all other predictors ($x_{j \neq i}$) fixed**

EX: "With a fixed radio budget, how much increased sales do we get per extra $1000 spent on TV?"

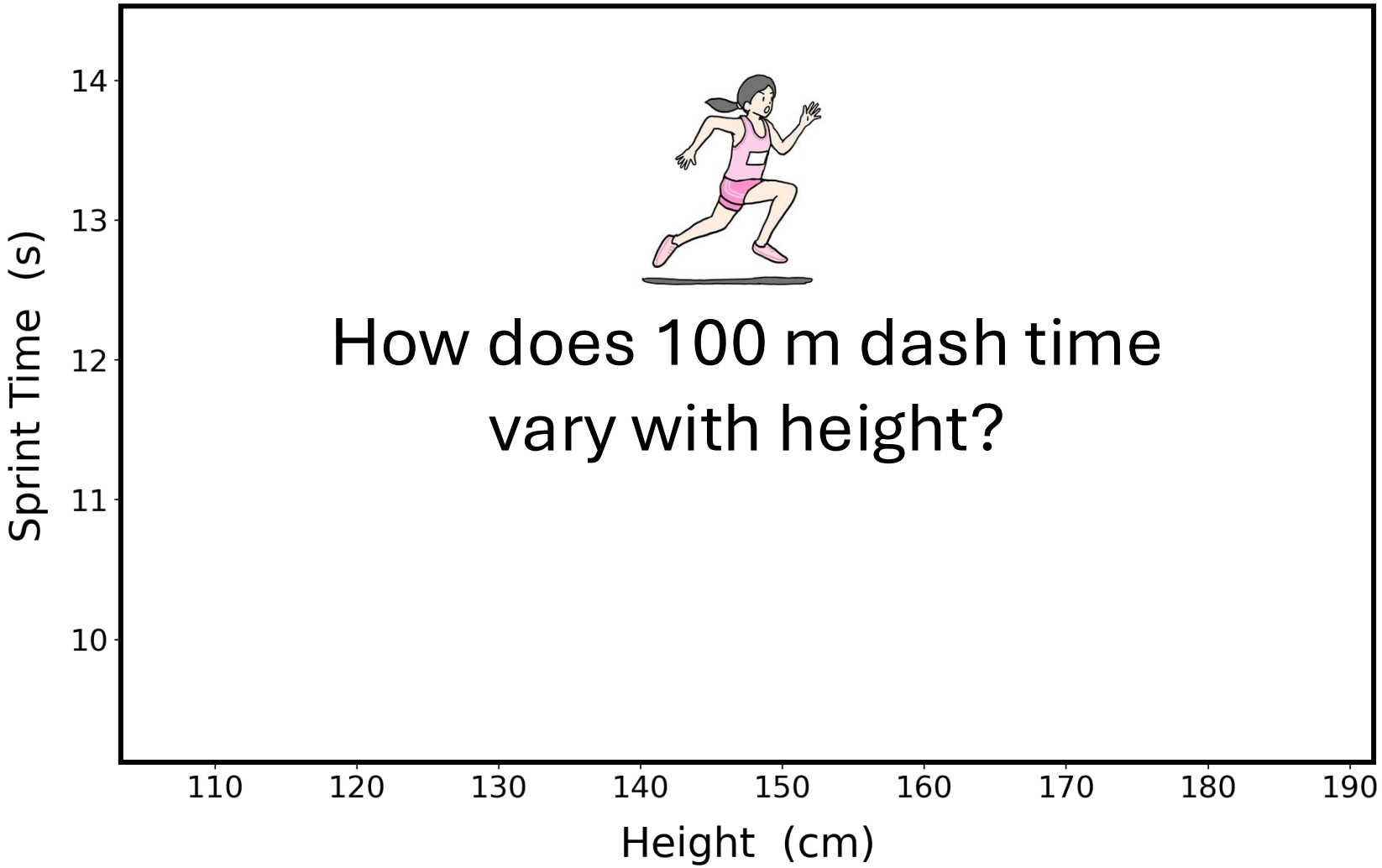
Each β_i represents a slope

These are "slices" of the plane



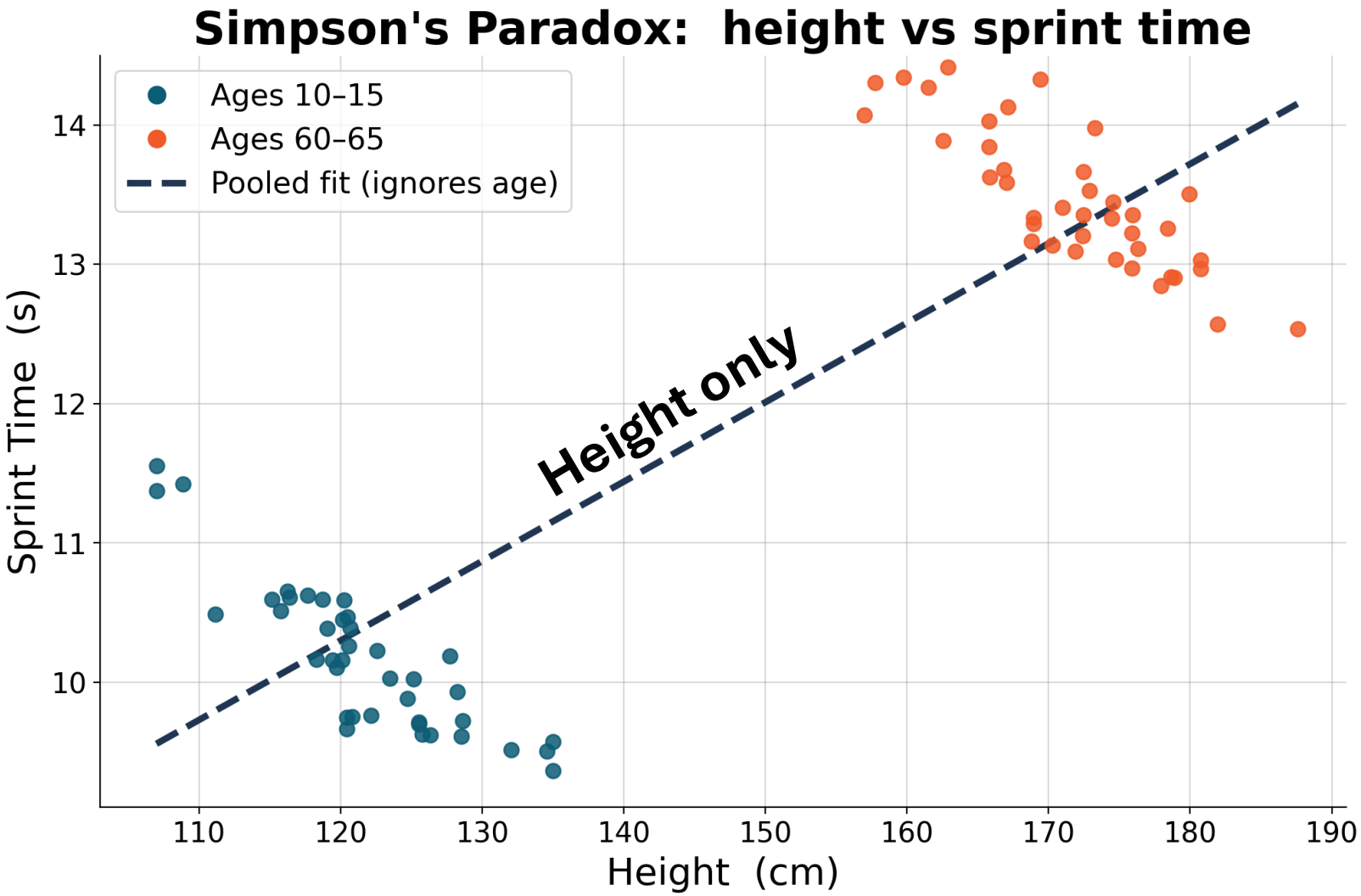


Holding other predictors fixed *vs.* ignoring them



Suppose we first try to model this using only height, **ignoring all other possible variables** that might affect running speed.

Holding other predictors fixed *vs.* ignoring them

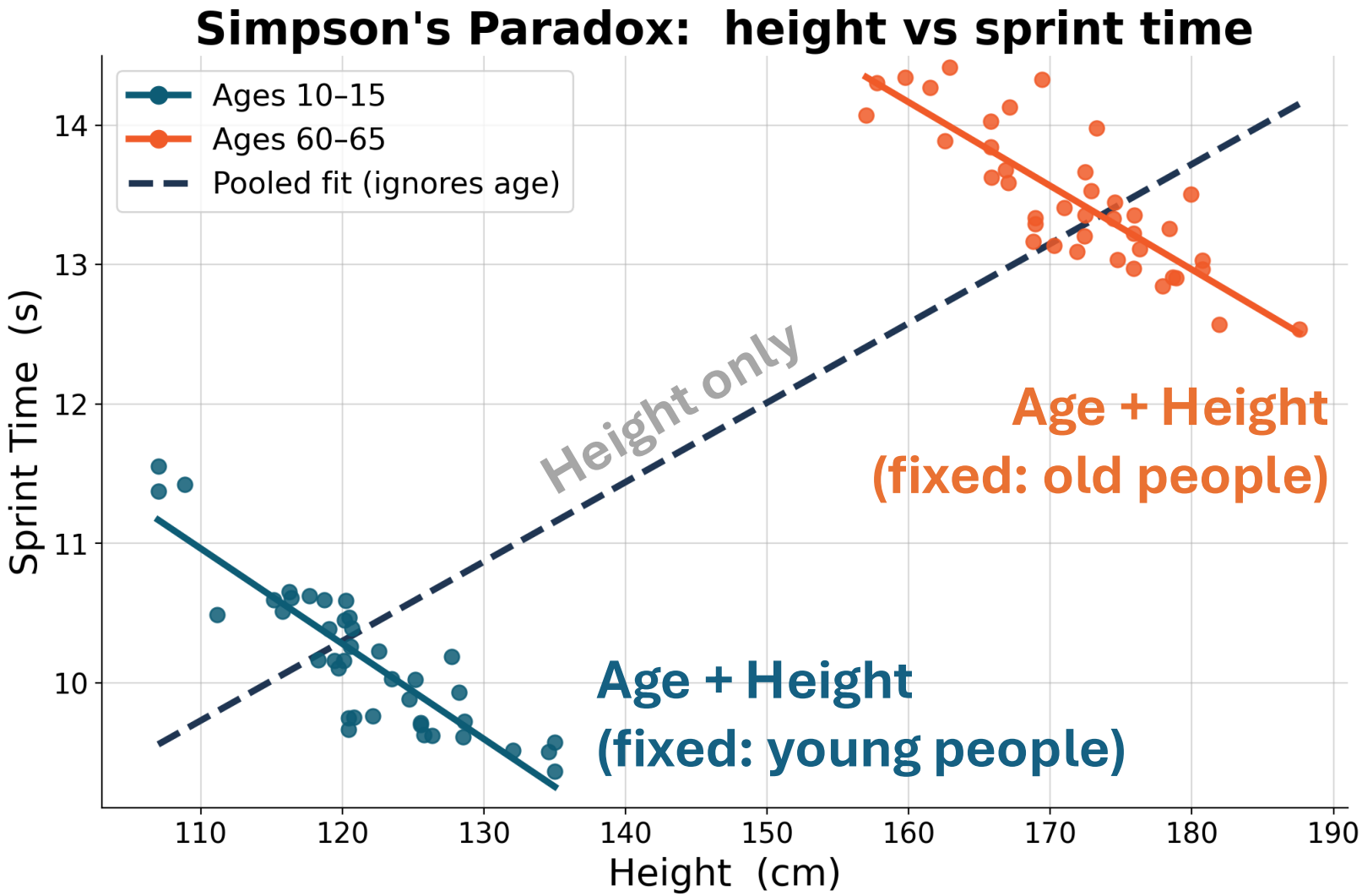


Using only age, the linear model tell us:

"Tall people run slower"

Incorrect!

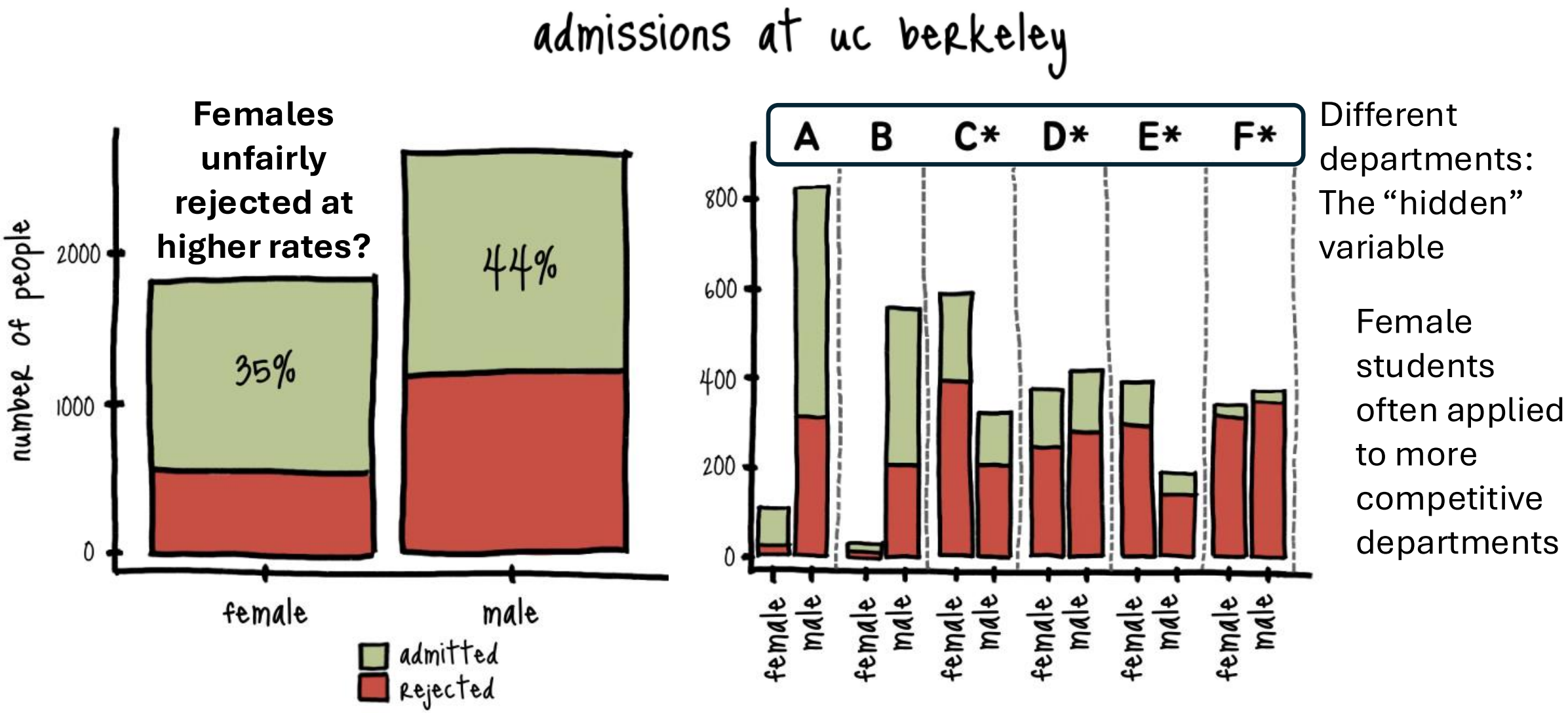
Holding other predictors fixed *vs.* ignoring them



Using only age, the linear model tell us:
"Tall people run slower"
Incorrect!

By doing *multiple* linear regression, our slopes effectively hold age fixed while modeling height:
"Short people run slower"
"Old people run slower"
"At fixed age, short people run slower" *Correct!*

An interesting example of Simpson's paradox



Different departments: The "hidden" variable

Female students often applied to more competitive departments

Visuals from a LinkedIn post by Maciej Zimoch

Back to the meaning of β_i: What does sign tell us?

$$t_{\text{sprint}} = \beta_0 + \beta_1(\text{height}) + \beta_2(\text{age})$$

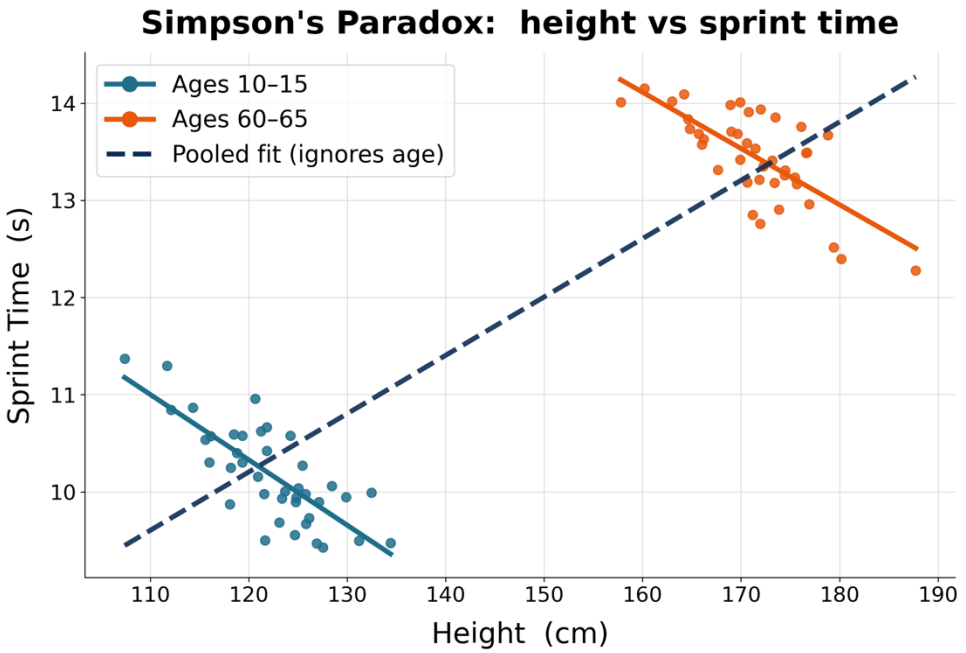
β_1 **will be negative** (faster sprint)

e.g., $\beta_1 = -0.05$ **sec/cm**

β_2 **will be positive** (slower sprint)

e.g., $\beta_2 = +0.12$ **sec/yr**

Always write units!!!



"Tall people run faster"
"Old people run slower"

Does $|\beta_2| > |\beta_1|$ mean age is "more important" than height when it comes to running?

Quantifying relative feature importance

β_1 will be negative (faster sprint)

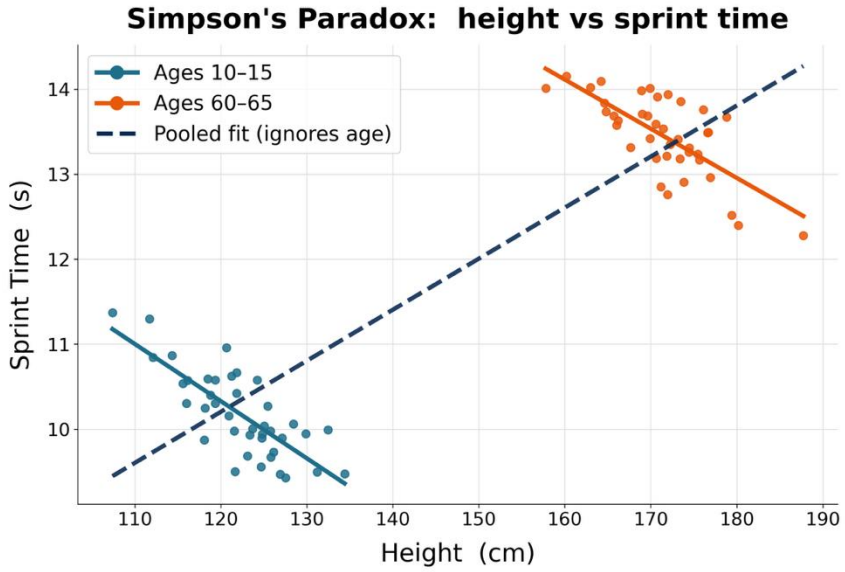
e.g., $\beta_1 = -0.05$ **sec/cm**

β_2 will be positive (slower sprint)

e.g., $\beta_2 = +0.12$ **sec/yr**

These are **different units**, and therefore, **different scales**

Cannot simply compare directly!



In our sprint time example, age and height span a reasonably similar(-ish) range of values:

$$105 \text{ cm} < \text{height} < 135 \text{ cm}$$
$$10 \text{ yrs} < \text{age} < 65 \text{ yrs}$$

But often times this will not be the case; features span different orders of magnitude

Recall: Features live on different scales



Suppose we are building a model to predict car price based on its **mileage** and **number of previous owners**



Number of previous owners: 1-5

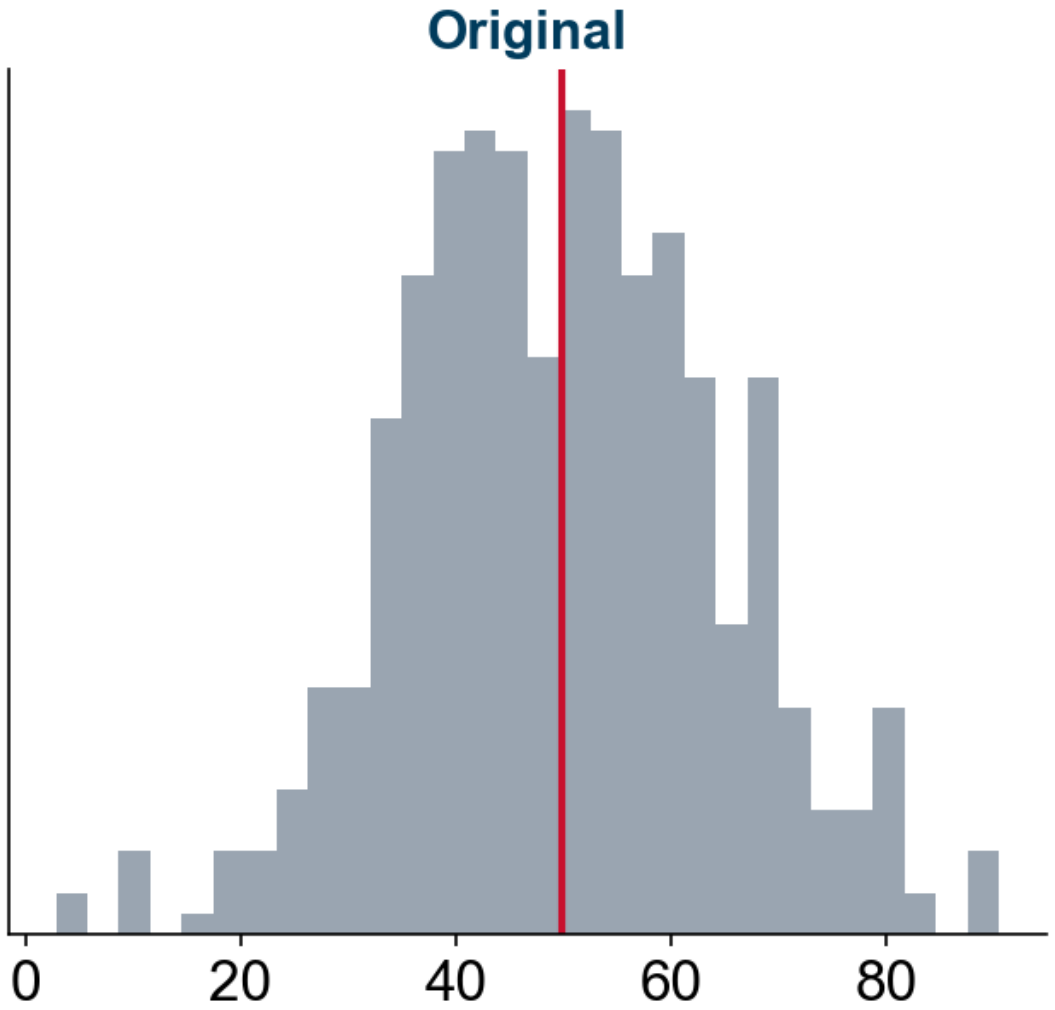


Mileage: anywhere from 0 to 300,000 miles



Without scaling, **mileage values are thousands of times larger** and can dominate distance-based analyses, even though previous owners may be just as influential.

Recall: Standardization (Z-score)



Each sample

Mean of the data

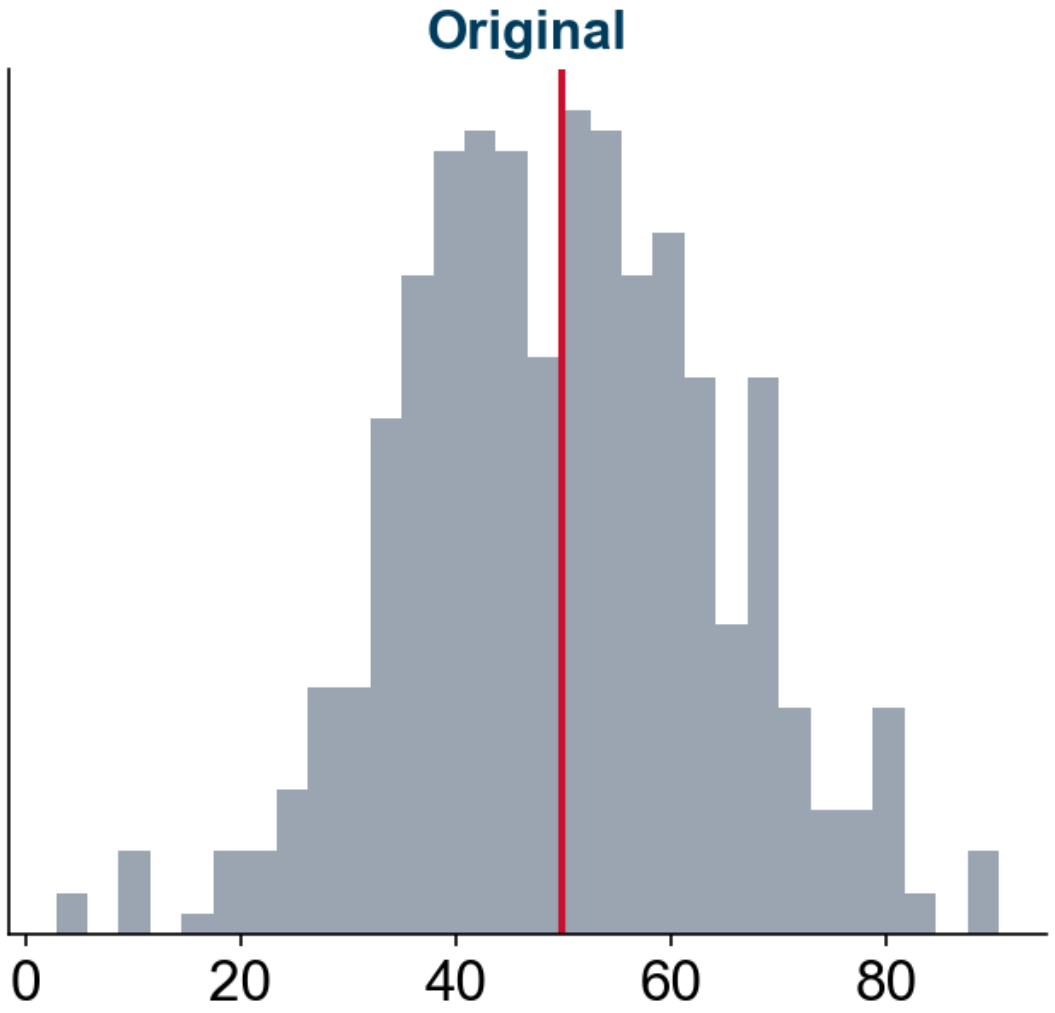
$$Z = \frac{x - \mu}{\sigma}$$

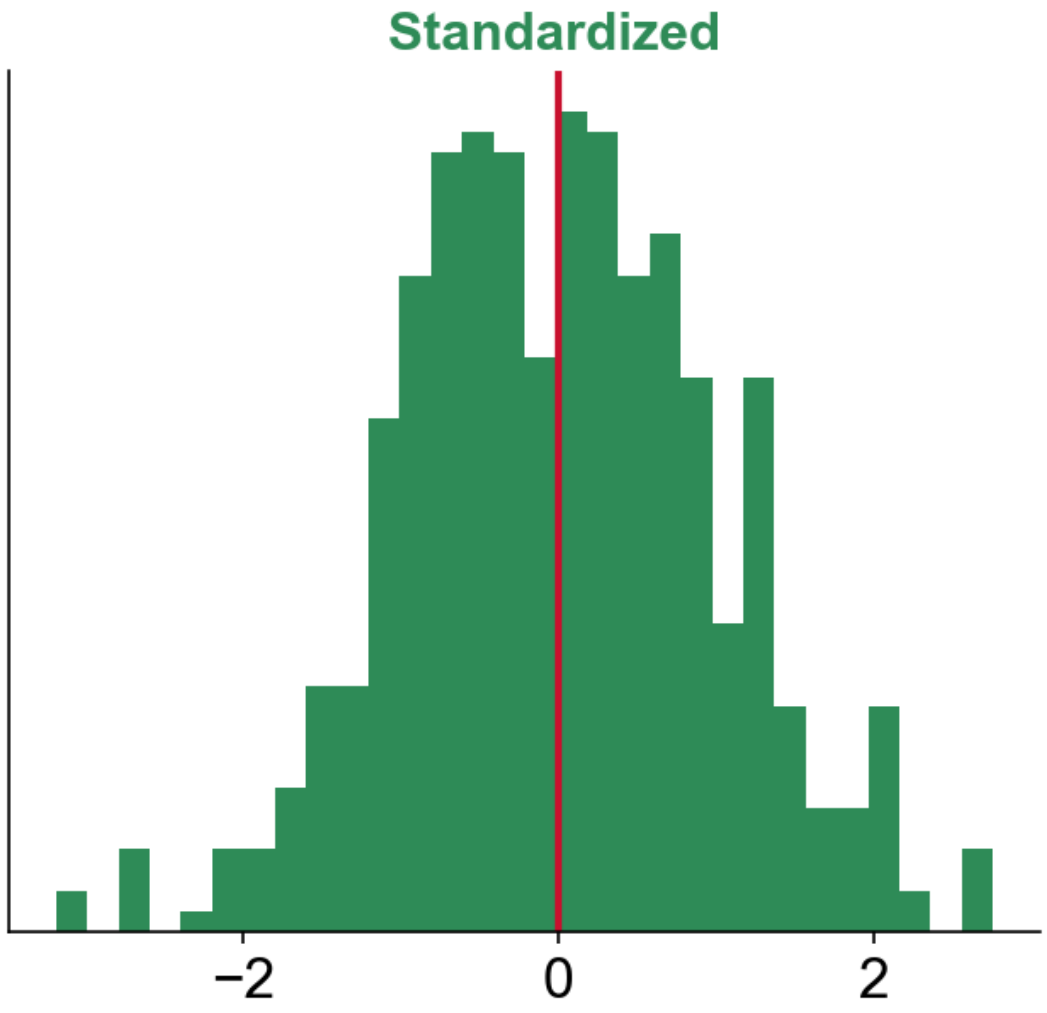
Standard deviation

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \sigma^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \mu)^2$$

1) Subtract the mean to **zero-center** data
2) Divide by Z to **standardize** it

Recall: Standardization (Z-score)





Effects of standardization on β_i

Number of owners	Mileage	Sale price
2	80k	$15k
1	65k	$20k
3	122k	$11k
2	150k	$10k
1	103k	$21k
1	21k	$28k
2	88k	$18k
3	182k	$6k
5	190k	$4k
4	159k	$8k

Without any standardization:

$\beta_{\text{owners}} = -\$1.6\text{k per owner}$

$\beta_{\text{mileage}} = -\$0.10\text{k per mile}$

→ Number of owners matters 16× more(?)

No!

***With Z*-score standardization:**

$\beta_{\text{owners}} = -\$2.0\text{k per std dev in owners}$

$\beta_{\text{mileage}} = -\$5.2\text{k per std in mileage}$

→ **Mileage matters more than num owners!**

Questions?

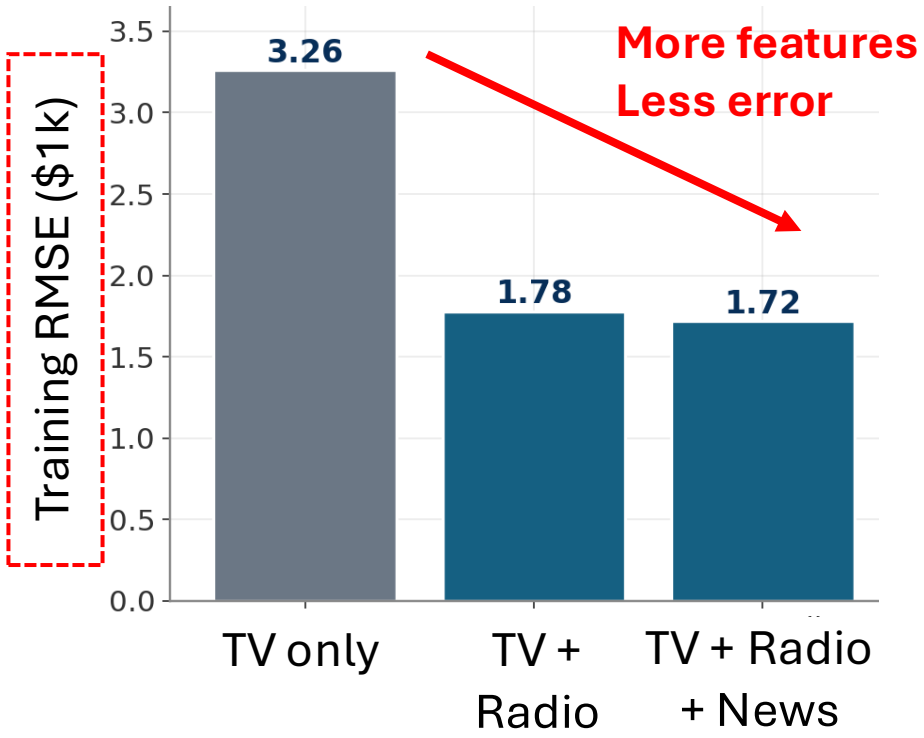
How many predictors to include?

Adding more predictors (usually) helps

Sales of a product depends on advertising through different venues



TV budget	Radio budget	Paper budget	Unit sales
$100k	$8k	$59k	11k
$80k	$25k	$18k	9k
$195k	$42k	$92k	18k
$232k	$21k	$5k	23k



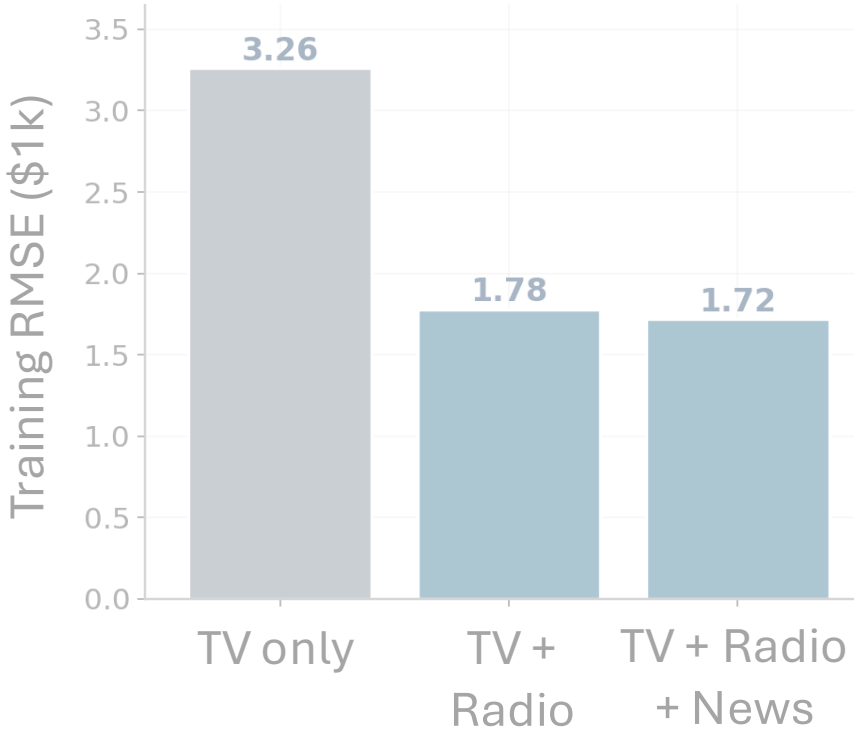
$$\hat{y} = \beta_0 + \beta_1 x_{\text{TV}} + \beta_2 x_{\text{Radio}} + \beta_3 x_{\text{News}}$$

- As we add features, our predictions get better
- Each *useful* predictor explains more of the variance
- Notice that **adding radio helps a lot**, whereas **adding newspaper helps little** – this gives some insight on how much each feature matters

Adding more predictors (usually) helps

$$\hat{y} = \beta_0 + \beta_1 x_{\text{TV}} + \beta_2 x_{\text{Radio}} + \beta_3 x_{\text{News}} + \beta_4 x_{\text{Salary}}$$

What if we add **a predictor that has no effect on the target?**
For example, **employee salaries** should not affect sales



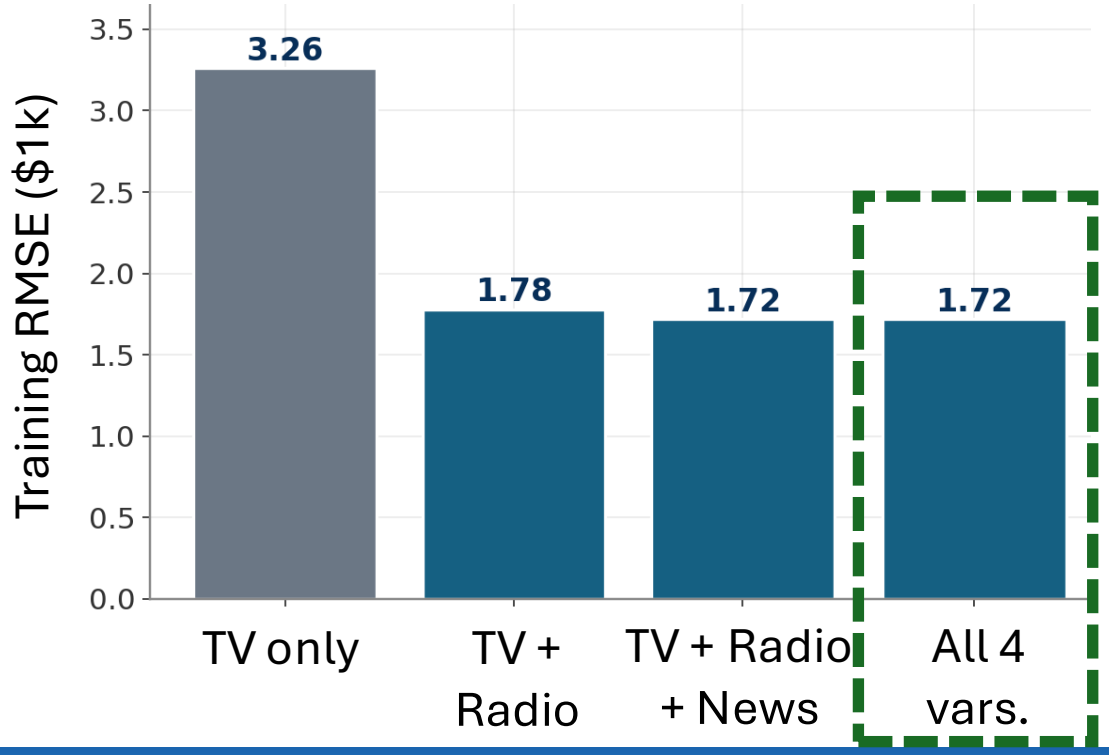
$$\hat{y} = \beta_0 + \beta_1 x_{\text{TV}} + \beta_2 x_{\text{Radio}} + \beta_3 x_{\text{News}}$$

- As we add features, our predictions get better
- Each *useful* predictor explains more of the variance
- Notice that **adding radio helps a lot**, whereas **adding newspaper helps little** – this gives some insight on how much each feature matters

…But not always!

$$\hat{y} = \beta_0 + \beta_1 x_{\text{TV}} + \beta_2 x_{\text{Radio}} + \beta_3 x_{\text{News}} + \beta_4 x_{\text{Salary}}$$

What if we add **a predictor that has no effect on the target?**
For example, **employee salaries** should not affect sales

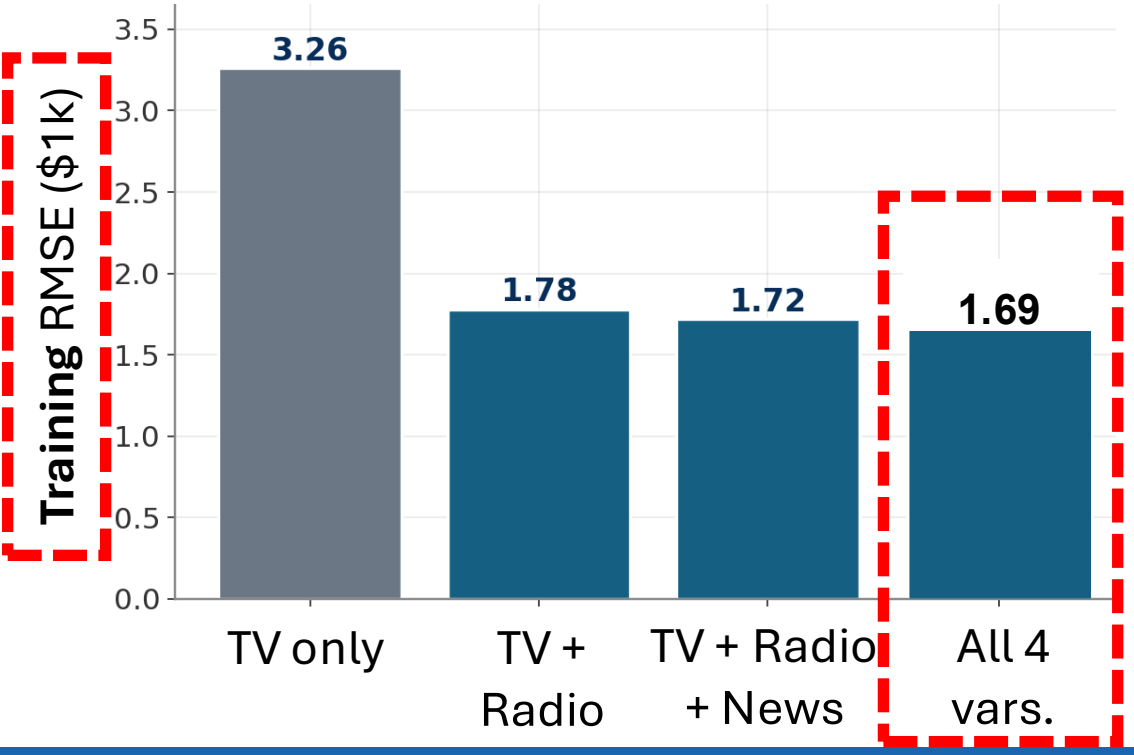


In the *best* case:
Our model will learn there is no correlation between sales and production cost ($\beta_4 = 0$), and therefore, **our predictions remain unaffected**

…But not always!

$$\hat{y} = \beta_0 + \beta_1 x_{\text{TV}} + \beta_2 x_{\text{Radio}} + \beta_3 x_{\text{News}} + \beta_4 x_{\text{Salary}}$$

What if we add **a predictor that has no effect on the target?**
For example, **employee salaries** should not affect sales



In the *worst* case:
Our model learns unphysical correlations between sales and production cost, leading to **lower RMSE on the *training* set – but worse predictions on the *test* set**

So…don't add unnecessary features!

Solution #1: Always use a test set!

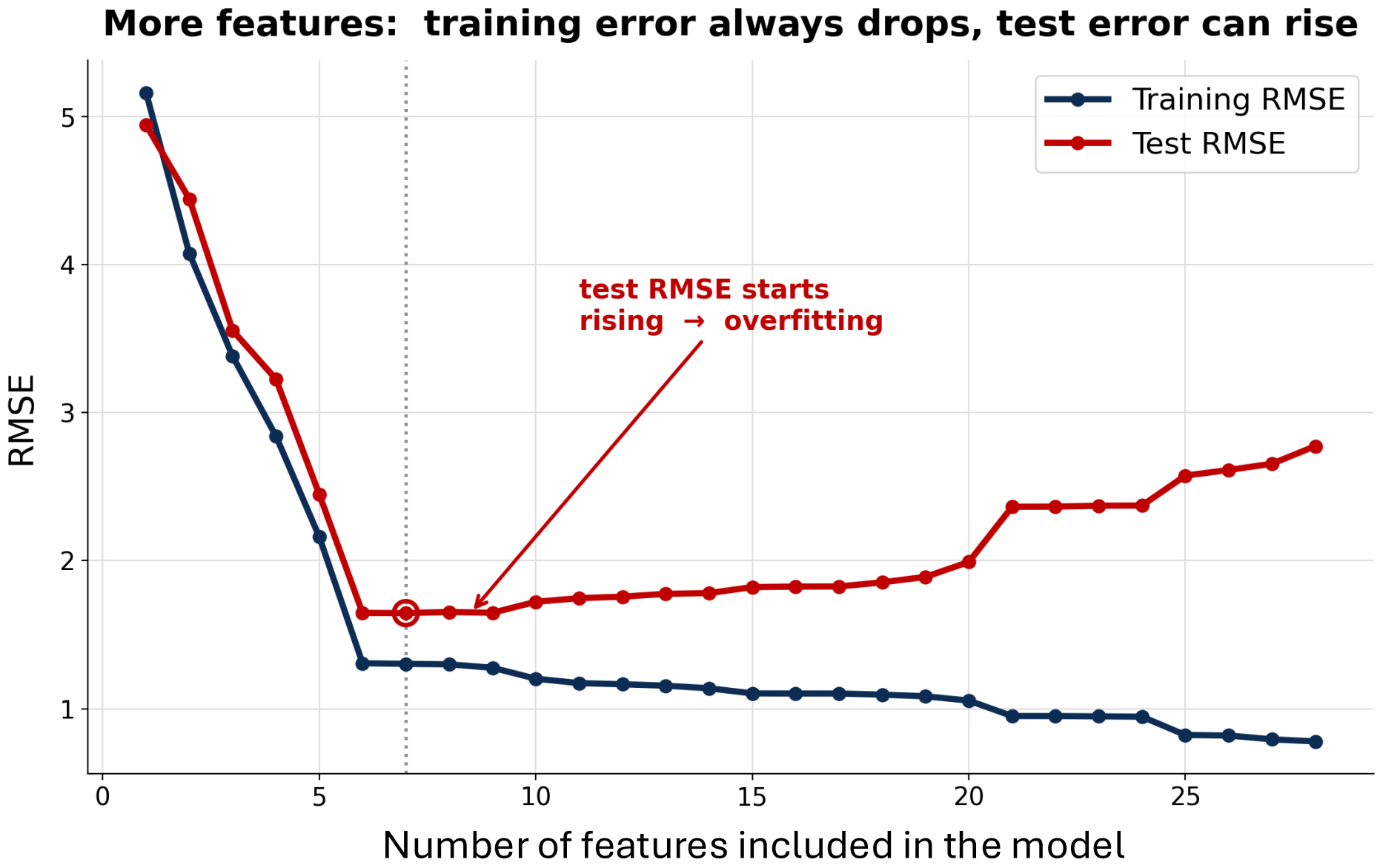


Model learns the task

How good is the model on "new" data?

Performance metrics are reported **on the test dataset**
These are the metrics we actually care about!

Solution #1: Always use a test set!



Solution #2: Report adjusted metrics

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

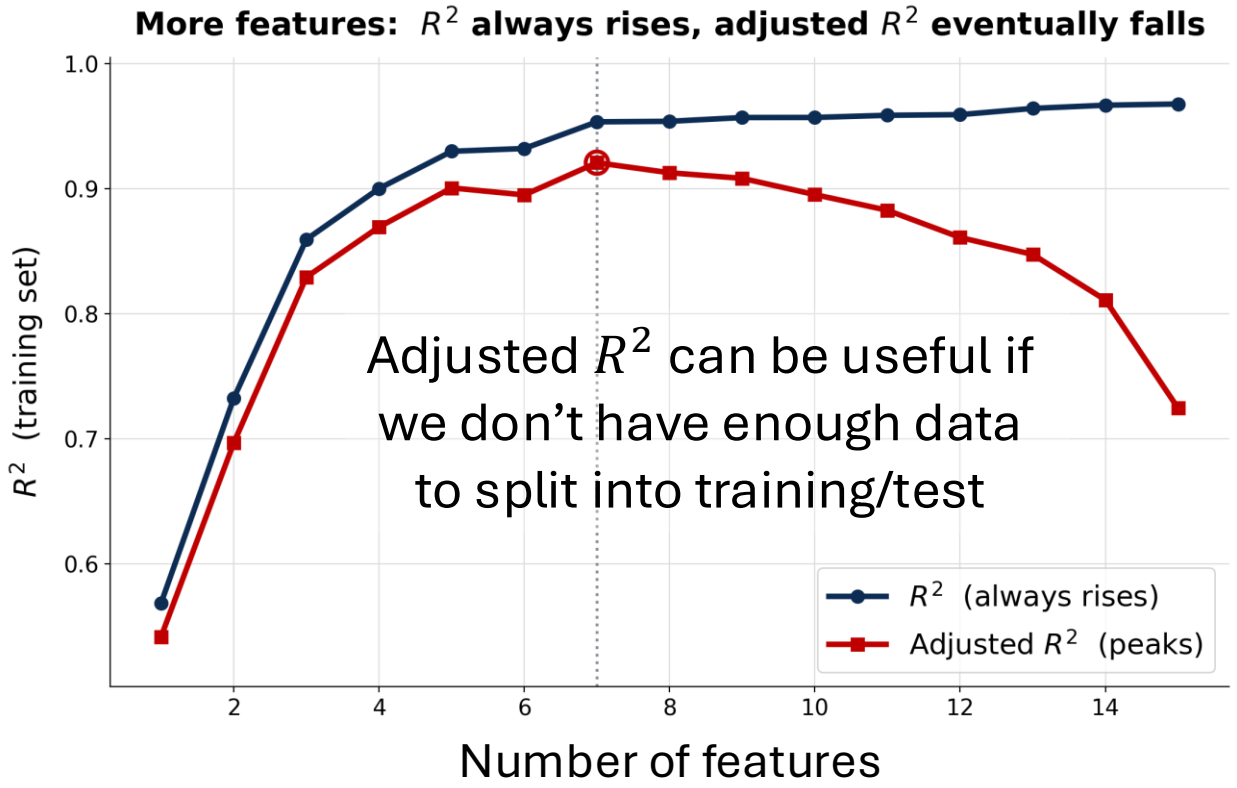
This will ***never*** **decrease** upon adding predictors

Num. Datapoints

$$R^2_{\text{adj}} = 1 - (1 - R^2)\frac{N - 1}{N - p - 1}$$

Num. Predictors

This will only increase if the new variable improves the model more than expected by chance



In general: $p < N - 1$

If $p = 0$, then $R^2_{\text{adj}} = R^2$

If p gets large while R^2 levels off, R^2_{adj} gets small

Another reason to avoid unnecessary predictors

$$\beta = (X^T X)^{-1} X^T y$$

$(X^T X)^{-1}$ must be invertible!

With many variables/features,
we may encounter is *collinearity*
(linear dependence)

Examples of collinearity

$$\beta = (X^T X)^{-1} X^T y$$

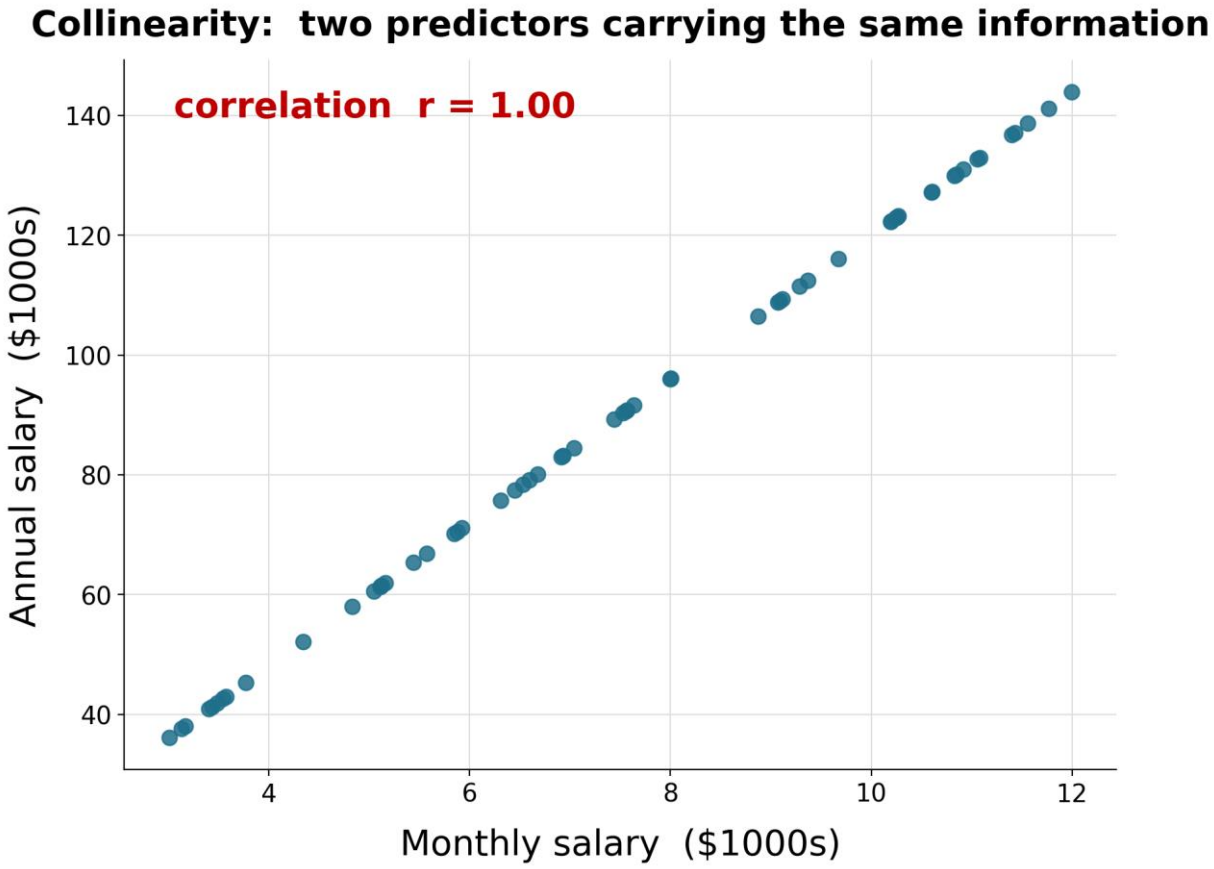
$(X^T X)^{-1}$ **must be invertible!**

Suppose a set of predictors includes:

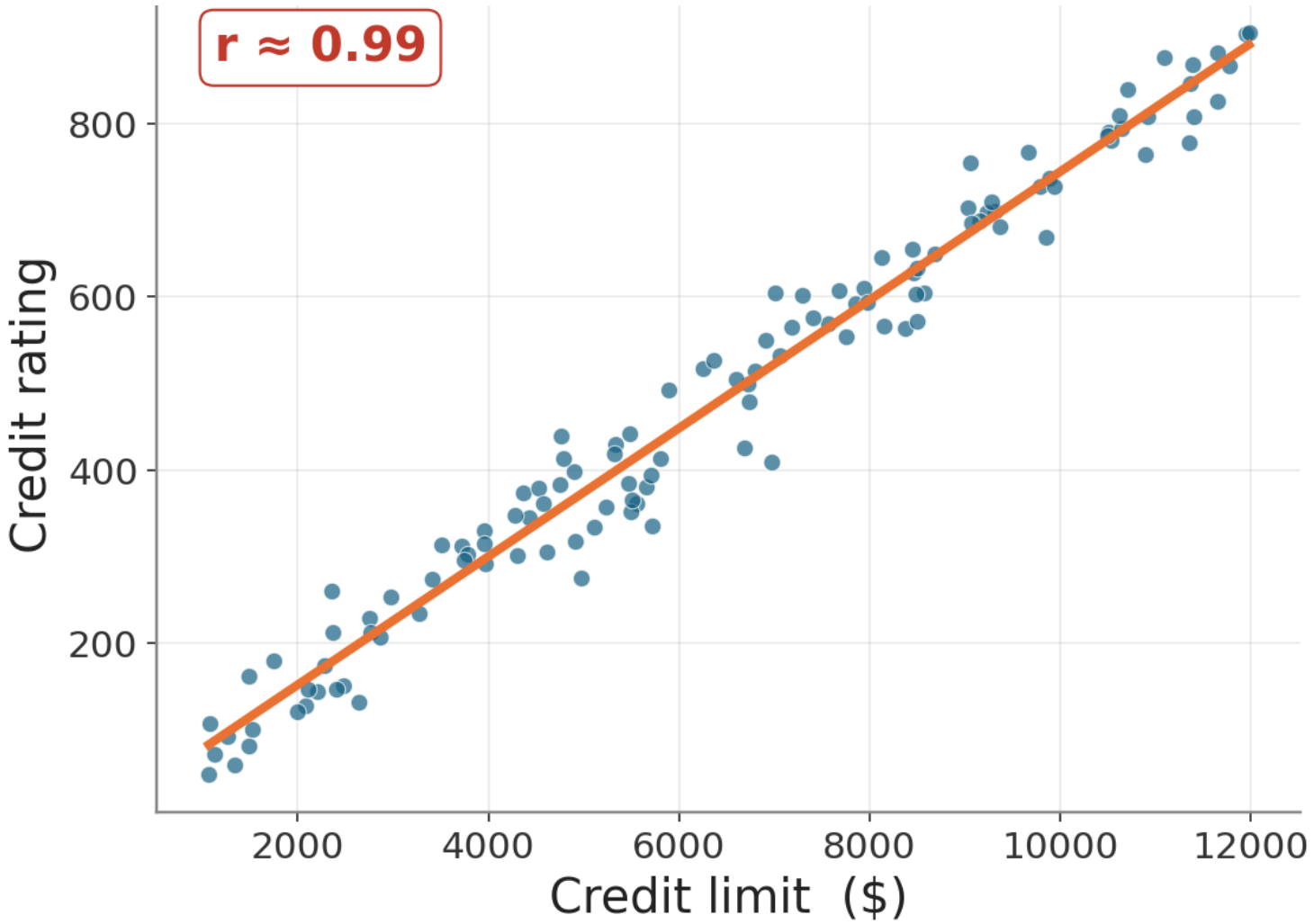
- Monthly salary ($)
- Annual salary ($)

These two variables are perfectly collinear – **one depends on the other!**

Our matrix inversion will fail



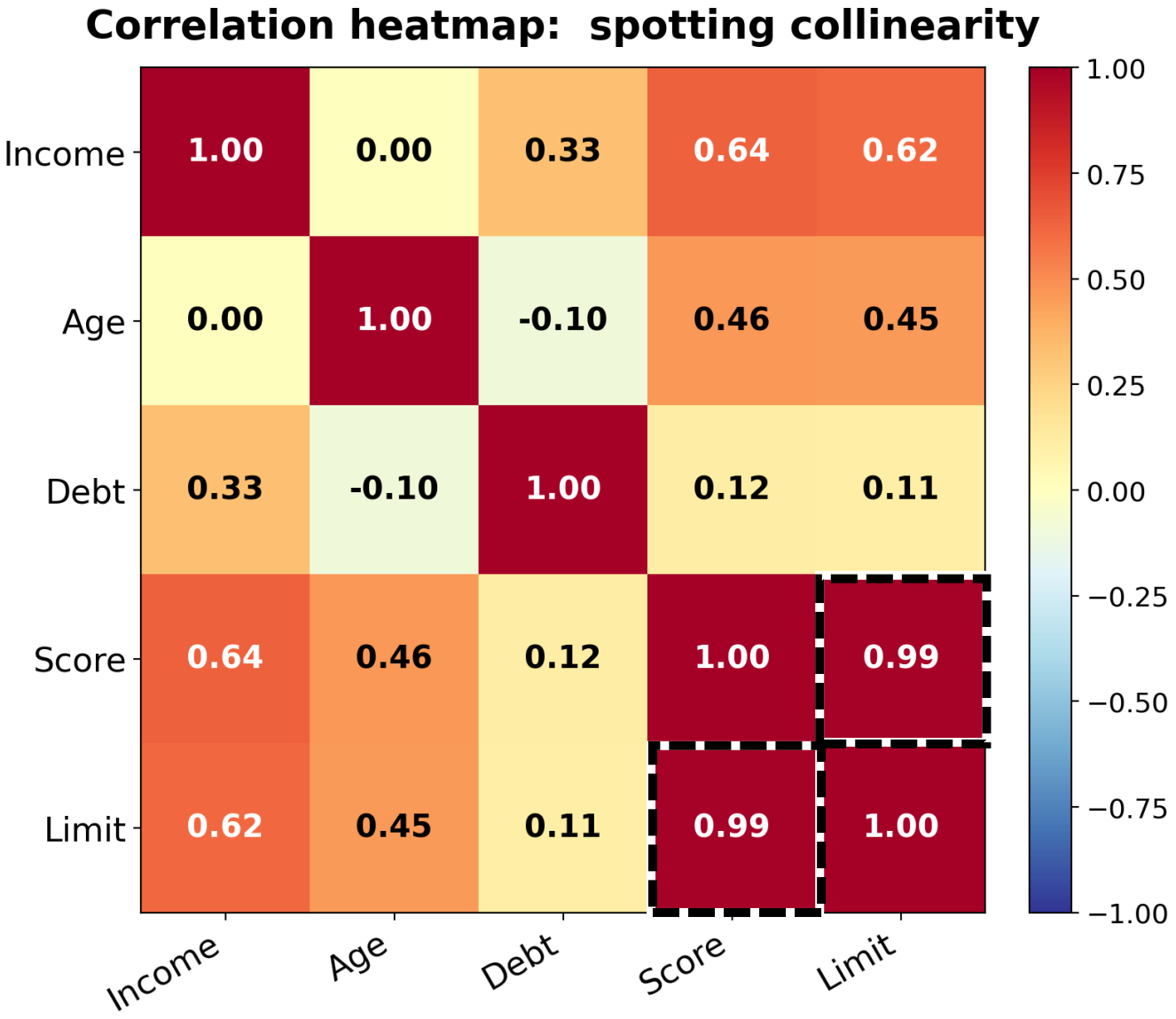
Examples of collinearity



Approximate collinearity (*e.g.*, credit limit and credit score) is more common, and still problematic…

- Highly unstable coefficients (the model is confused about which feature matters)
  $\rightarrow \beta_i$ fluctuate wildly
- Large standard error $SE(\beta_i)$

Another reason to always visualize your data!



Each value represents R^2 between the two variables

Off-diagonal components with values close to 1 are red-flags of collinearity!

Only keep one feature from the collinear pair

Questions?

Interaction Terms

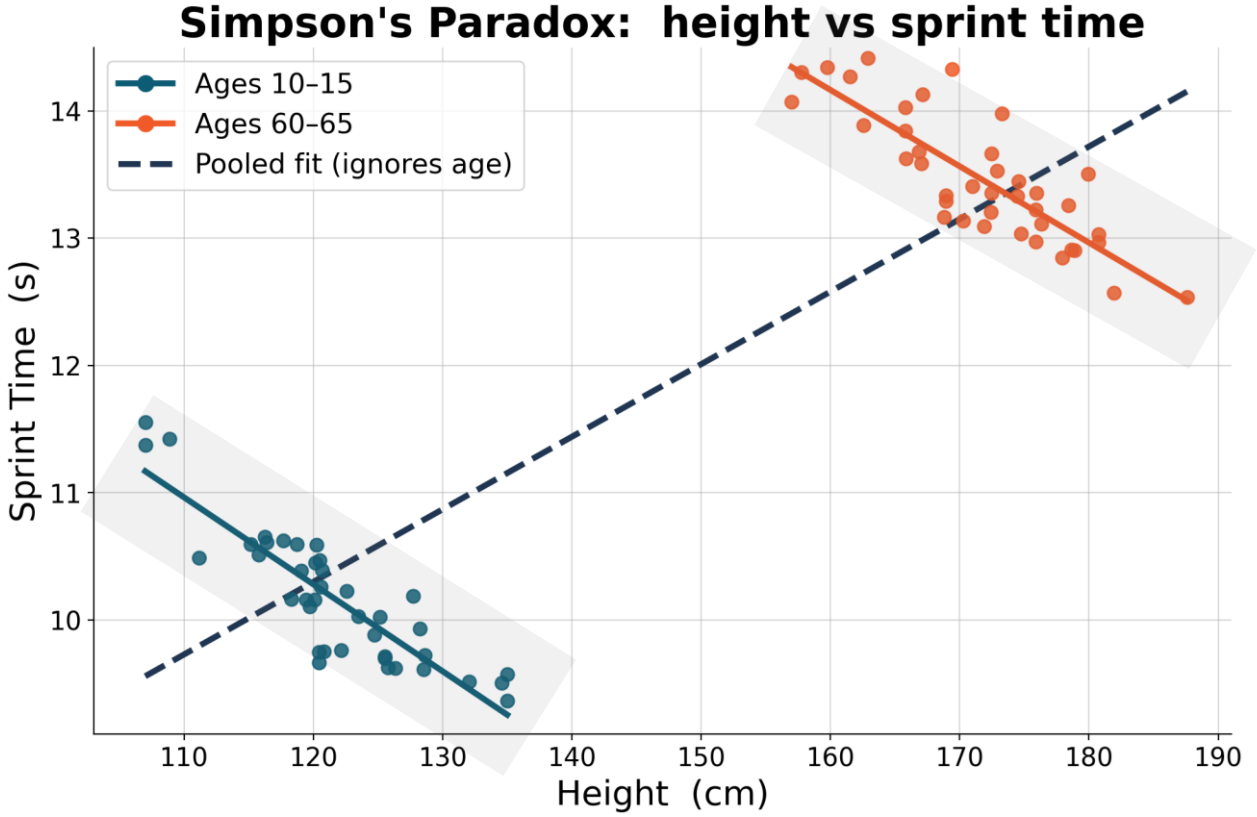
What if the value of one feature changes the effect of another?

Variables can sometimes *interact*

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

This assume x_1 and x_2 **act independently.**

Remember: β_1 represents the change in $\hat{y}$ per one-unit increase in x_1, **while holding x_2 fixed**



In this example, the slope: $time/height$ Is the same for old and young people.

But what if this were not the case?

Maybe for old people, height matters less than it does for young people?

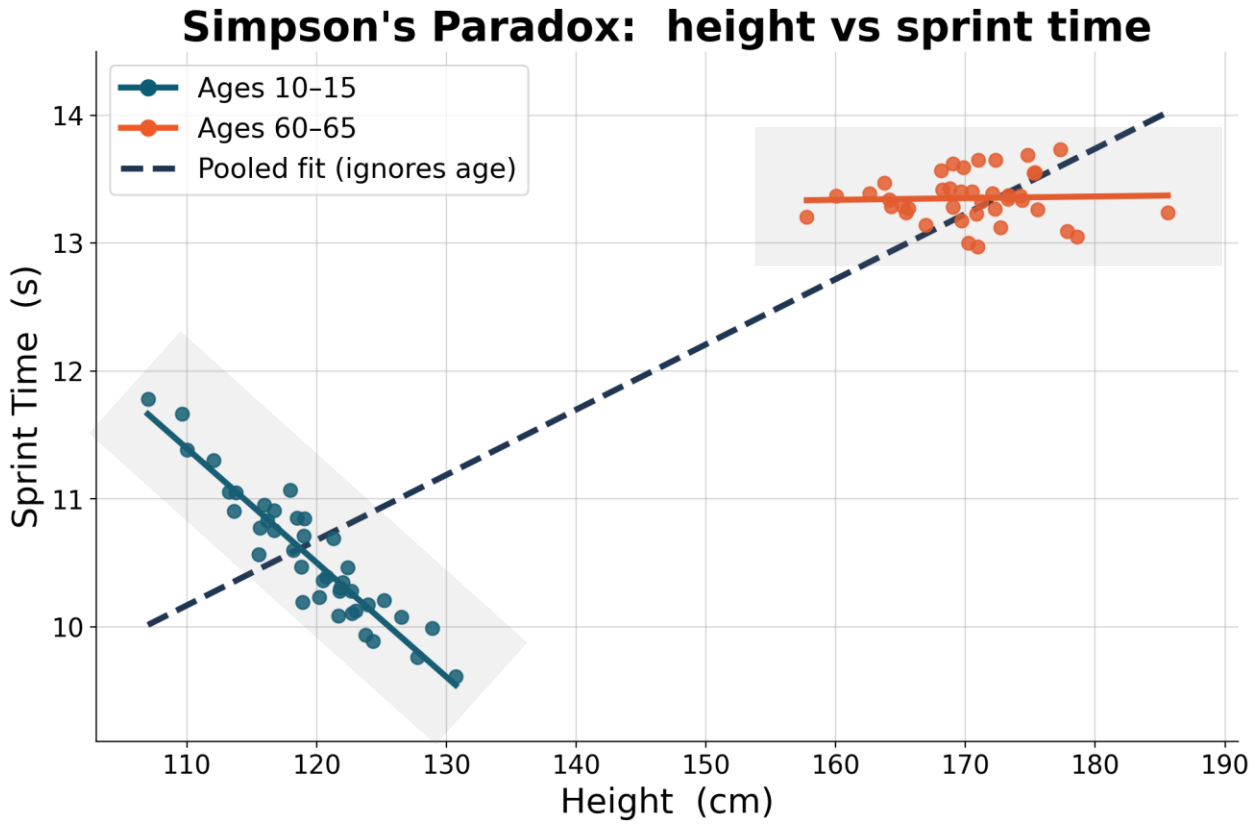
Variables can sometimes *interact*

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

This assume x_1 and x_2 **act independently.**

Remember: β_1 represents the change in $\hat{y}$ per one-unit increase in x_1, **while holding x_2 fixed**

Our current model cannot account for this! We need **interaction terms**



Some more examples of variable interactions

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

This assume x_1 and x_2 **act independently.**

Remember: β_1 represents the change in $\hat{y}$ per one-unit increase in x_1, **while holding x_2 fixed**

EX: Suppose we predicting **exam score** from…

- Hours studied
- Hours slept



What if the amount you studied only helps if you've gotten enough sleep?
In other words:

$\beta_{\text{hrs studied}}$ **should vary with** $x_{\text{hrs slept}}$

Variables can sometimes *interact*

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

This assume x_1 and x_2 **act independently.**

Remember: β_1 represents the change in $\hat{y}$ per one-unit increase in x_1, **while holding x_2 fixed**

EX: Suppose we predicting **exam score** from…

- Hours studied
- Hours slept



What if the amount you studied only helps if you've gotten enough sleep? In other words:

$\beta_{\text{hrs studied}}$ **should vary with** $x_{\text{hrs slept}}$

EX: Suppose we predicting **crop yield** from…



- Amount of water
- Amount of fertilizer

What if fertilizer only becomes effective after enough water has been added? In other words:

$\beta_{\text{fertilizer}}$ **should vary with** x_{water}

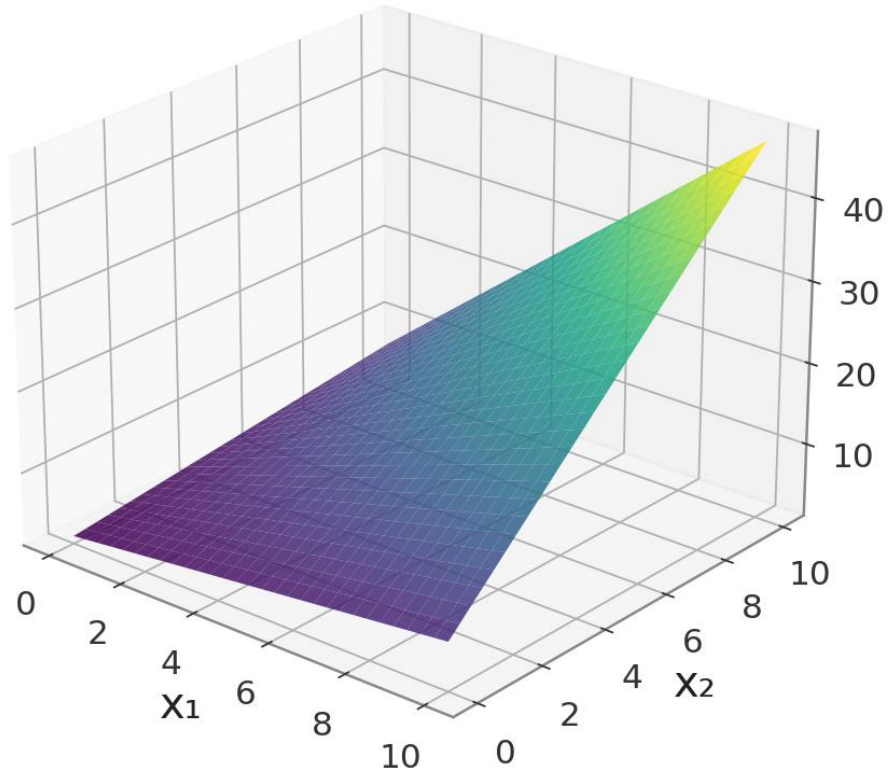
We account for this using *interaction terms*

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

This added terms let's the **predictors amplify or dampen one another**

Visually, it *twists* the plane

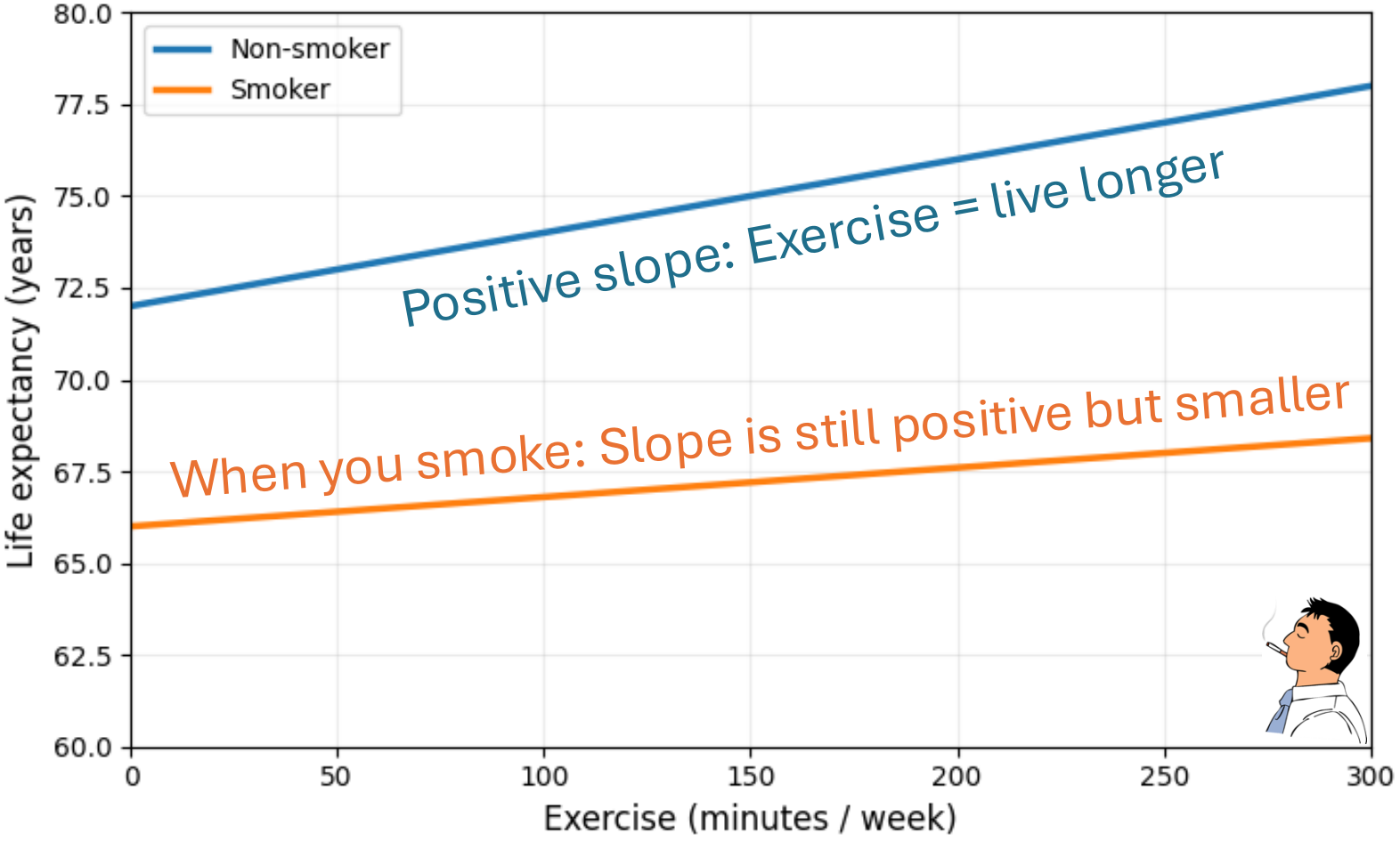
Slope changes with position along each axis



- $\beta_3 > 0 \rightarrow$ **effects amplify**
  - For example, water amplifies the effect of the fertilizer on crop yield, and sleep amplifies studying
- $\beta_3 < 0 \rightarrow$ **effects dampen**
  - For example, aging reduces the benefit of exercise on running speed

Another example, with visuals

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$



Suppose:

x_1 = hours exercised

x_2 = hours smoked

Based on the plot:

$\beta_3 < 0$

Exercise helps less when you smoke more

We account for this using *interaction terms*

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

This added terms lets the **predictors amplify or dampen one another**

How do we find the new coefficients?

$\hat{y}$ is still linear in β_i (in other words: we never multiple β_i by another β_j) so we can use the same math!

To illustrate this, suppose we set:

$$x_3 = x_1 x_2$$

- $\beta_3 > 0 \rightarrow$ **effects amplify**
  - For example, water amplifies the effect of the fertilizer on crop yield, and sleep amplifies studying
- $\beta_3 < 0 \rightarrow$ **effects dampen**
  - For example, aging reduces the benefit of exercise on running speed

We account for this using *interaction terms*

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

How do we find the new coefficients?

$\hat{y}$ is still linear in β_i (in other words: we never multiple β_i by another β_j) so we can use the same math!

To illustrate this, suppose we set:

$$x_3 = x_1 x_2$$

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

Then we solve by:

$$\frac{d\mathcal{L}}{d\beta_0} = 0, \frac{d\mathcal{L}}{d\beta_1} = 0, \frac{d\mathcal{L}}{d\beta_2} = 0, \frac{d\mathcal{L}}{d\beta_3} = 0$$

These derivatives don't care that $x_3 = x_1 x_2$, so we just use the same normal equation!

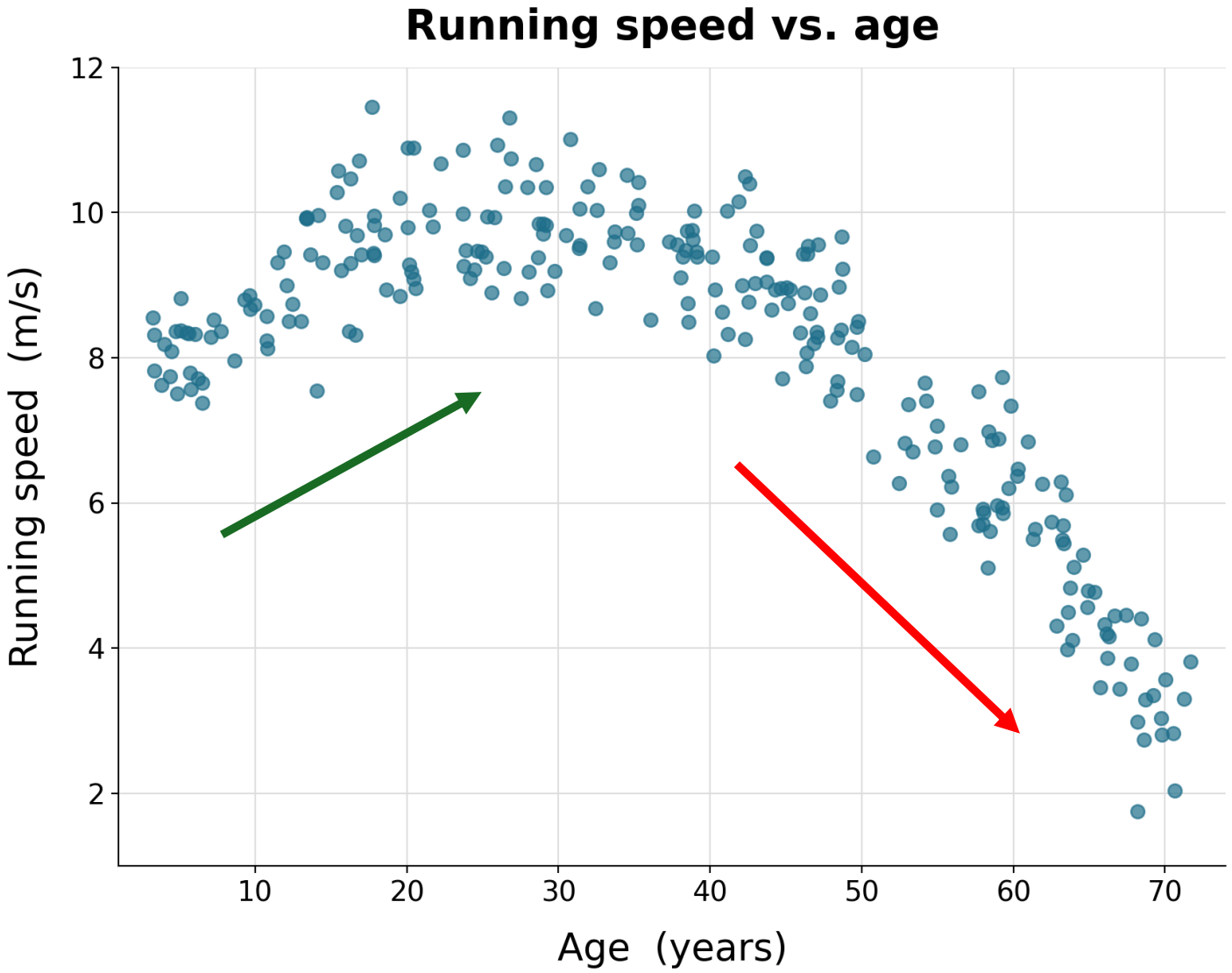
Questions?

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_1$$

Polynomial Regression

When variables interact with themselves

When are polynomials needed?

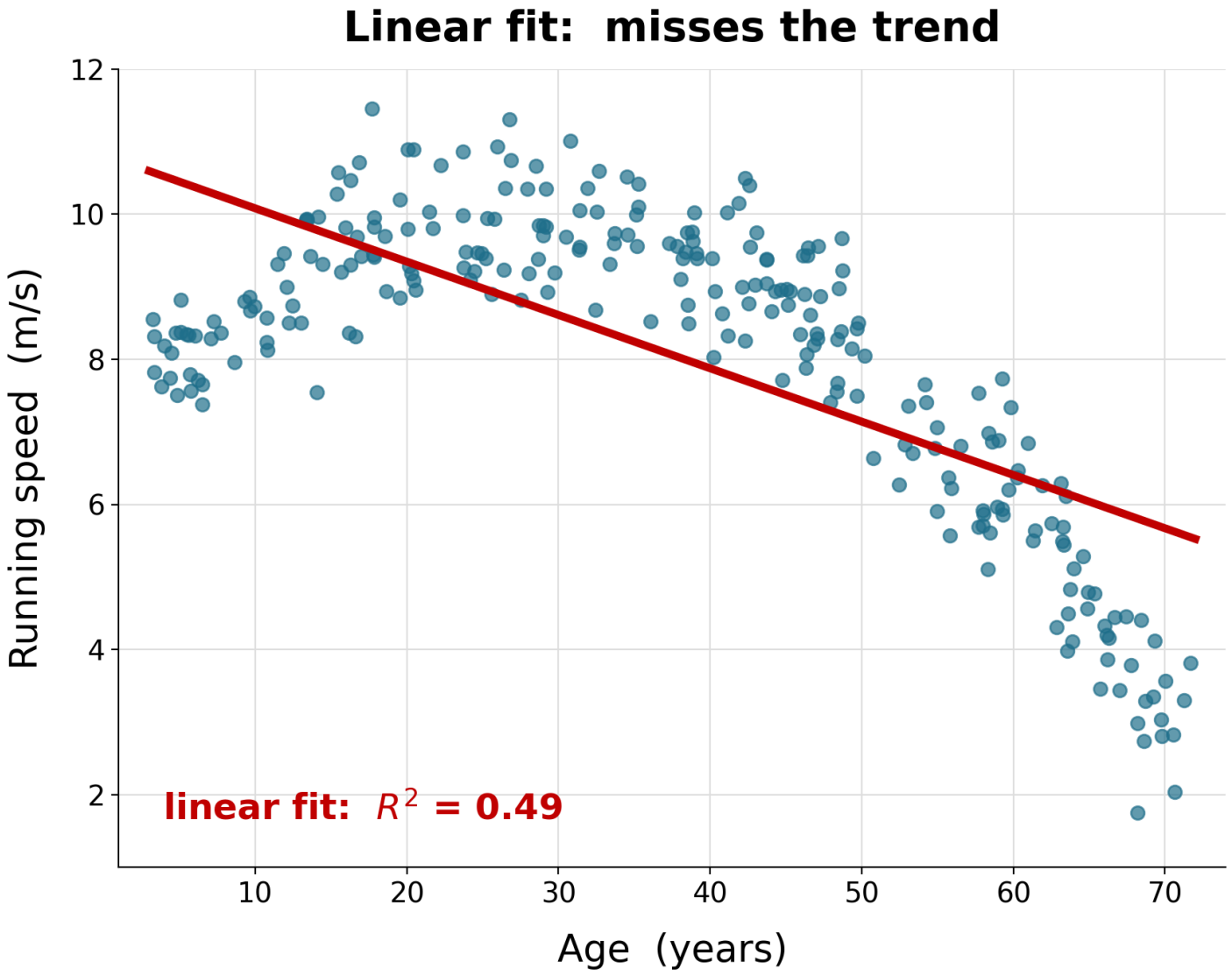


When there is some natural curvature in the data



- Children improve rapidly as they age.
- Adults eventually peak.
- Older adults begin to slow down.

When are polynomials needed?



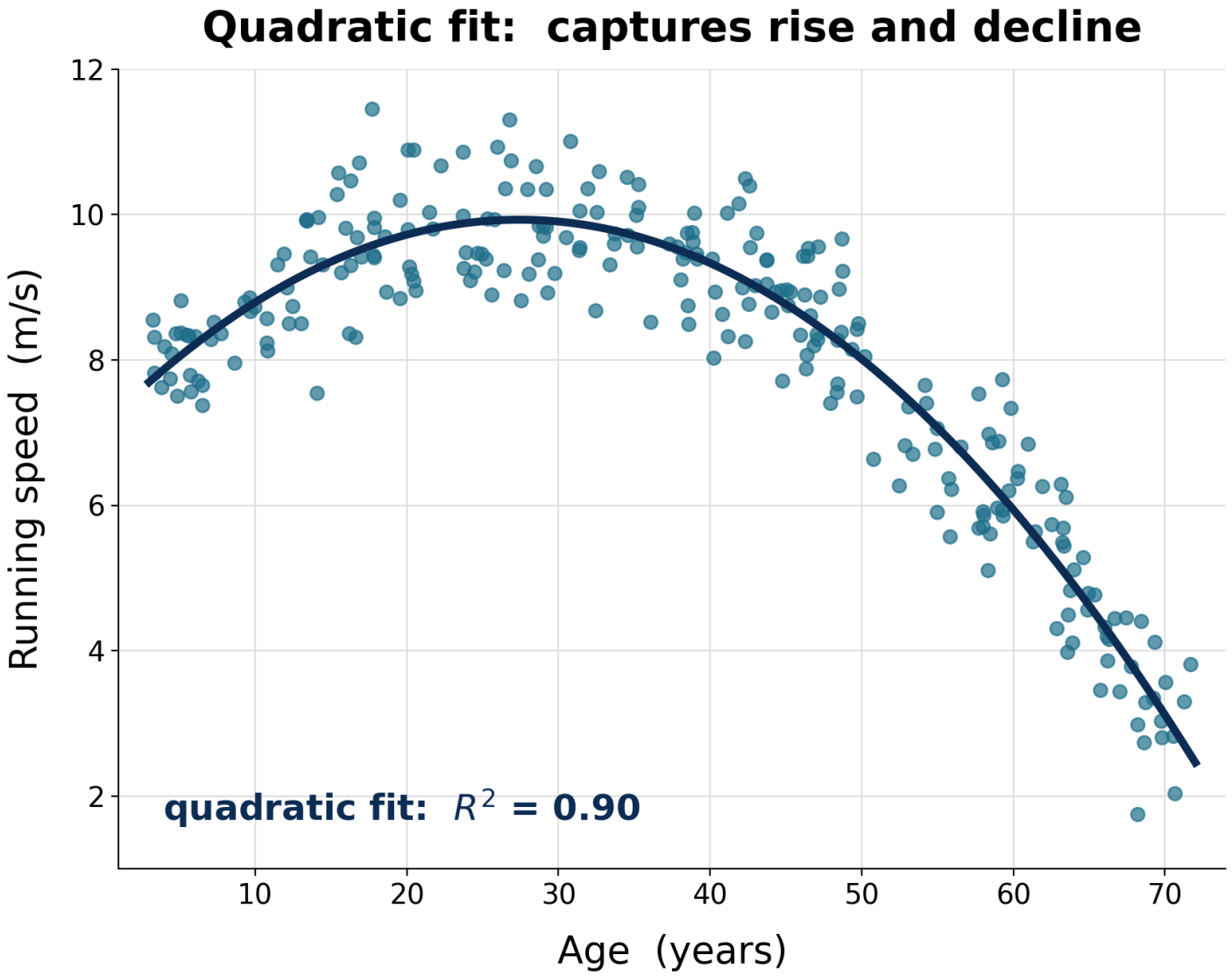
$$\hat{y}_{\text{speed}} = \beta_0 + \beta_1 x_{\text{age}}$$

This implicitly assumes a linear relationship.

Each year added leads to a fixed increase/decrease in running speed…

***Not* a good model**

When are polynomials needed?



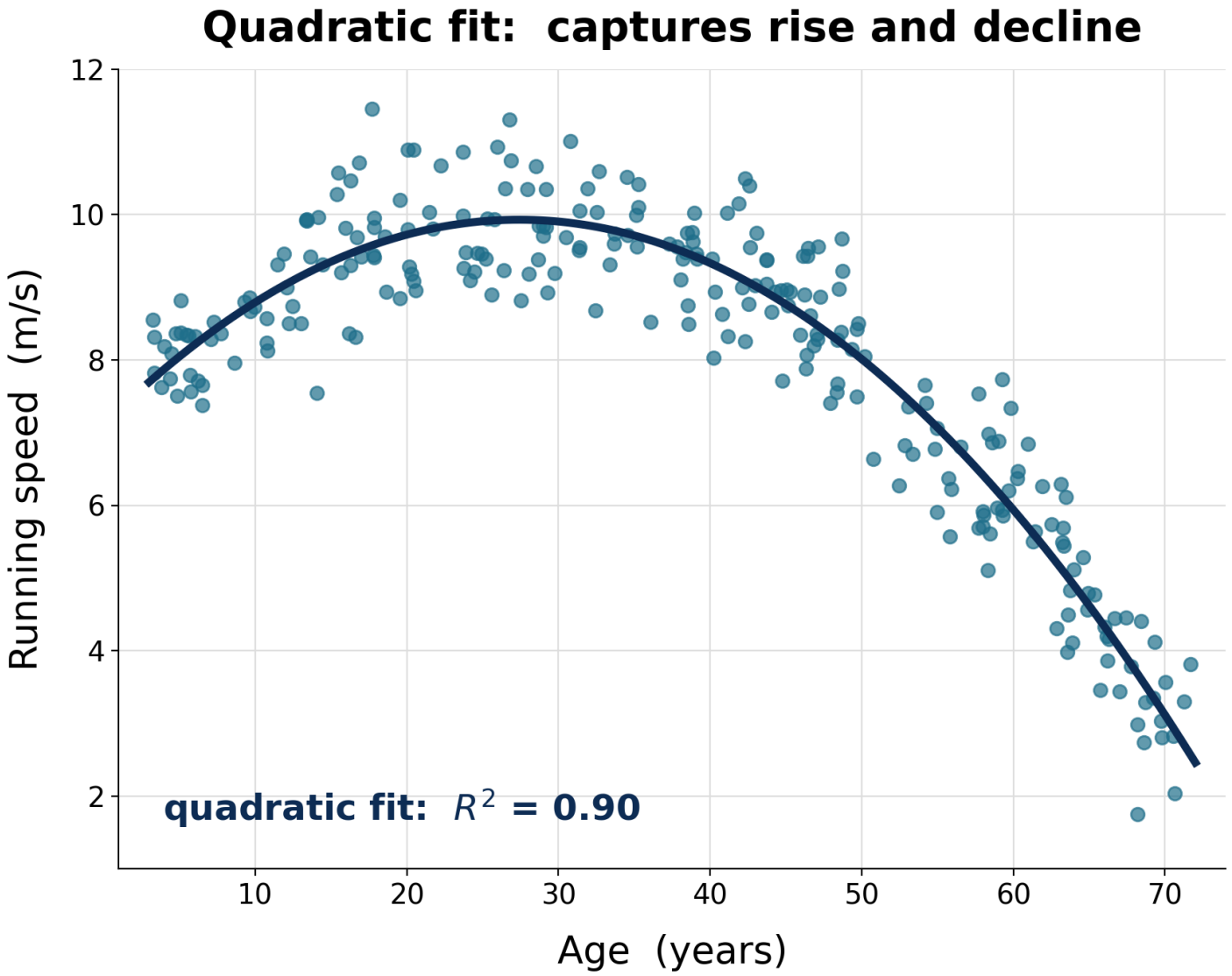
$$\hat{y}_{\text{speed}} = \beta_0 + \beta_1 x_{\text{age}} + \beta_2 x_{\text{age}}^2$$

The effect of age on speed depends on age itself!

Result: a much better fit!

And importantly, $\hat{y}$ **is still linear in** β_i, so our same normal equation applies

When are polynomials needed?

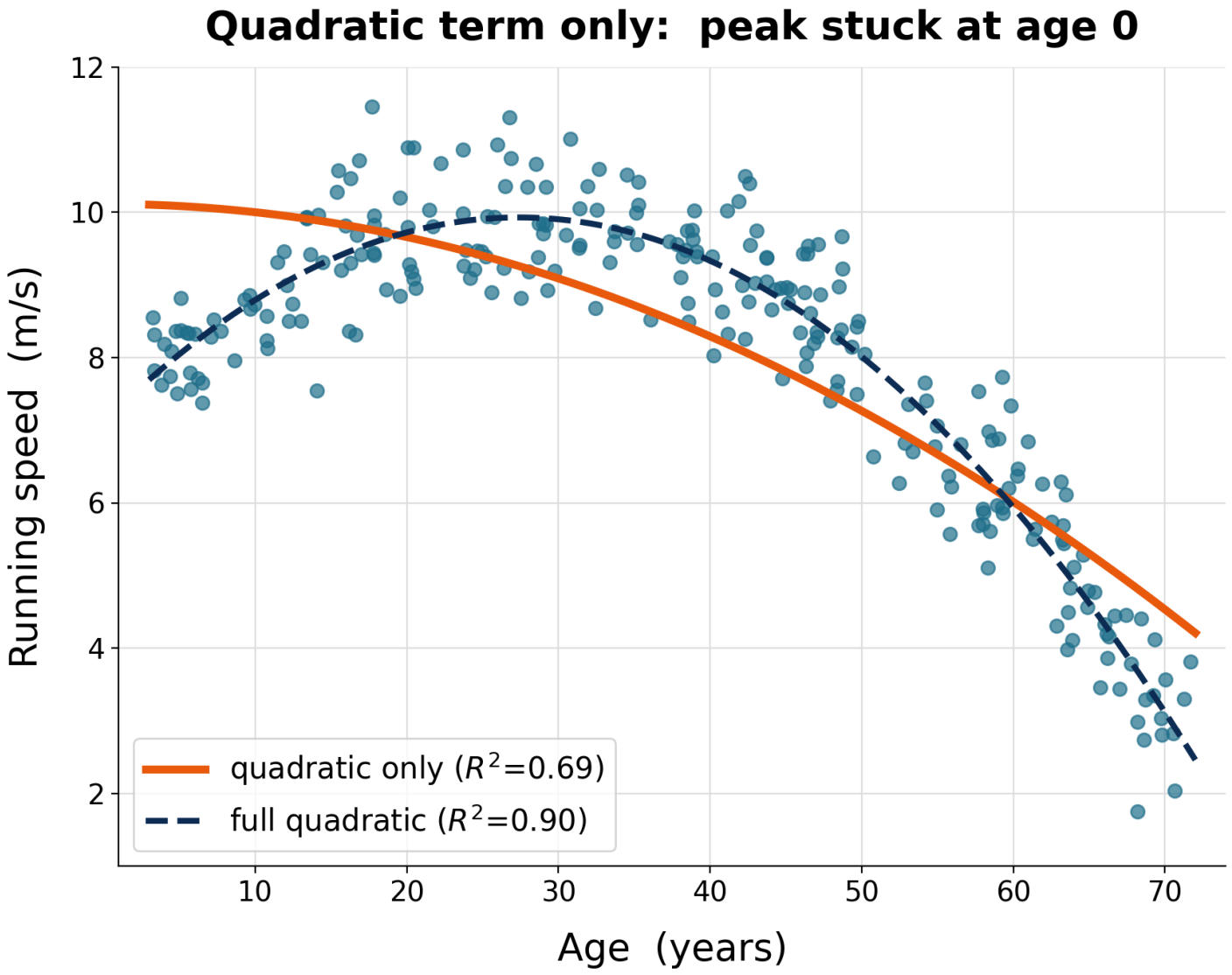


$$\hat{y}_{\text{speed}} = \beta_0 + \beta_1 x_{\text{age}} + \beta_2 x_{\text{age}}^2$$

β_1 decides where the vertex sits

β_2 controls the direction of the parabola and its curvature

When are polynomials needed?



$$\hat{y}_{\text{speed}} = \beta_0 + \beta_1 x_{\text{age}} + \beta_2 x_{\text{age}}^2$$

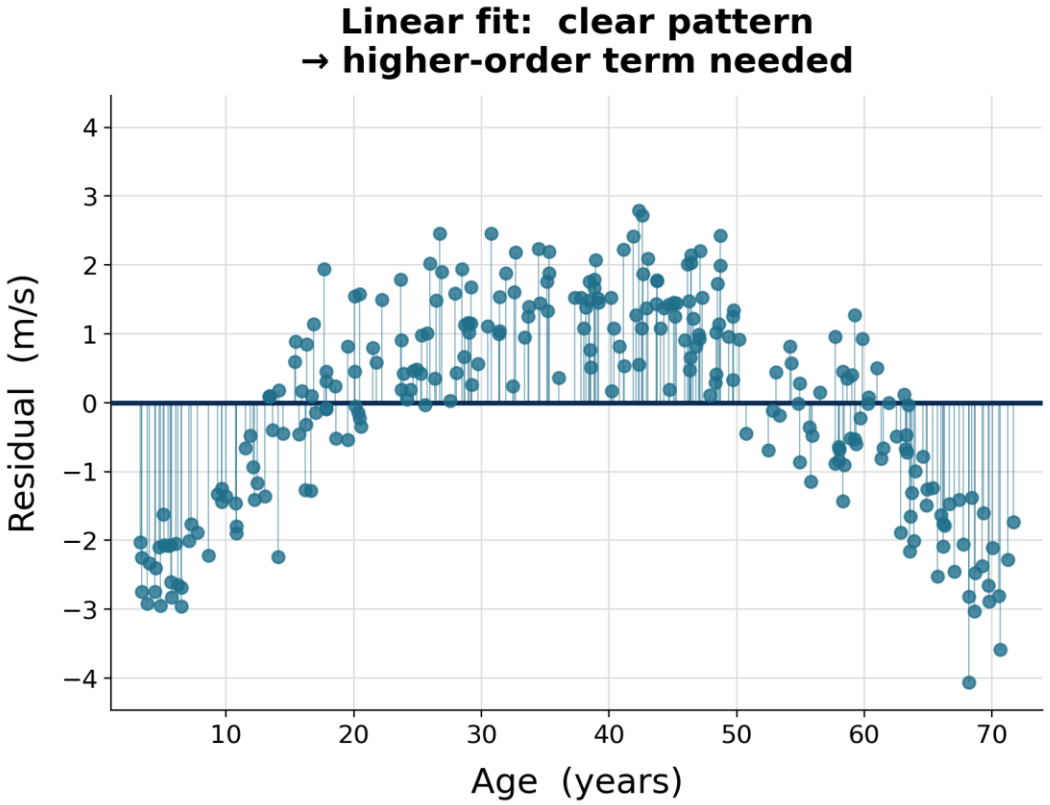
Typically, we need these lower-order terms, unless your data is symmetric around zero

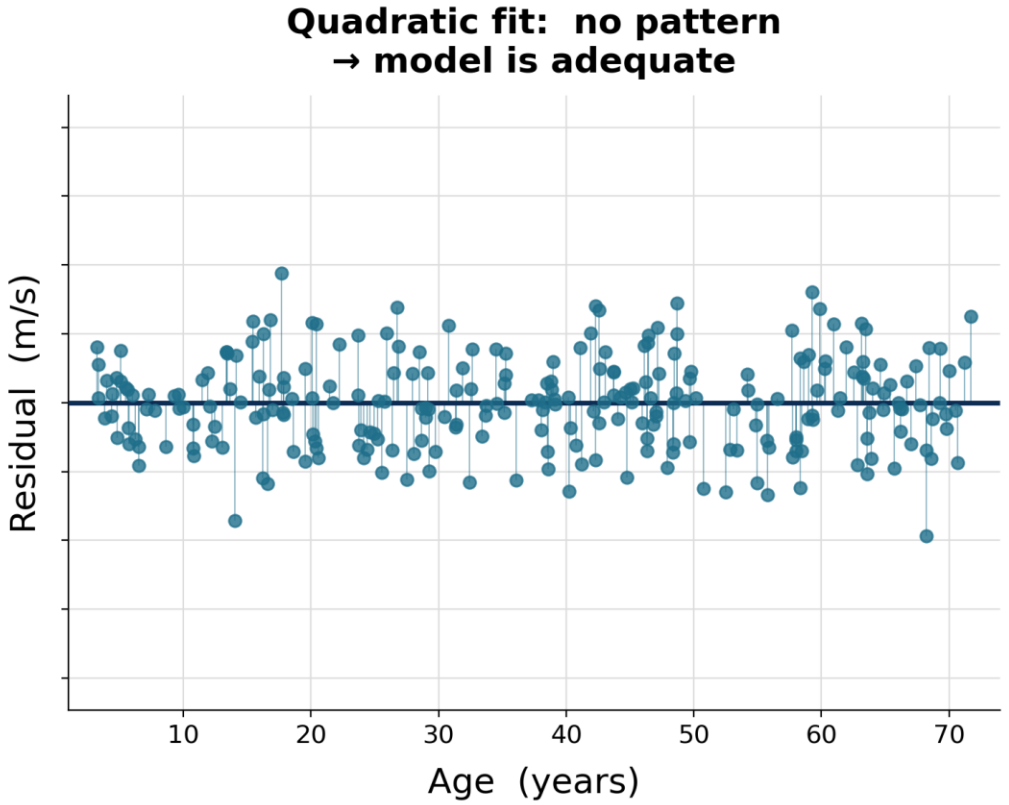
Generalizing to higher-order terms

$$\hat{y} = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \beta_n x^n$$

We can add arbitrarily higher orders – think of these as their own features.

But how do we know enough is enough?

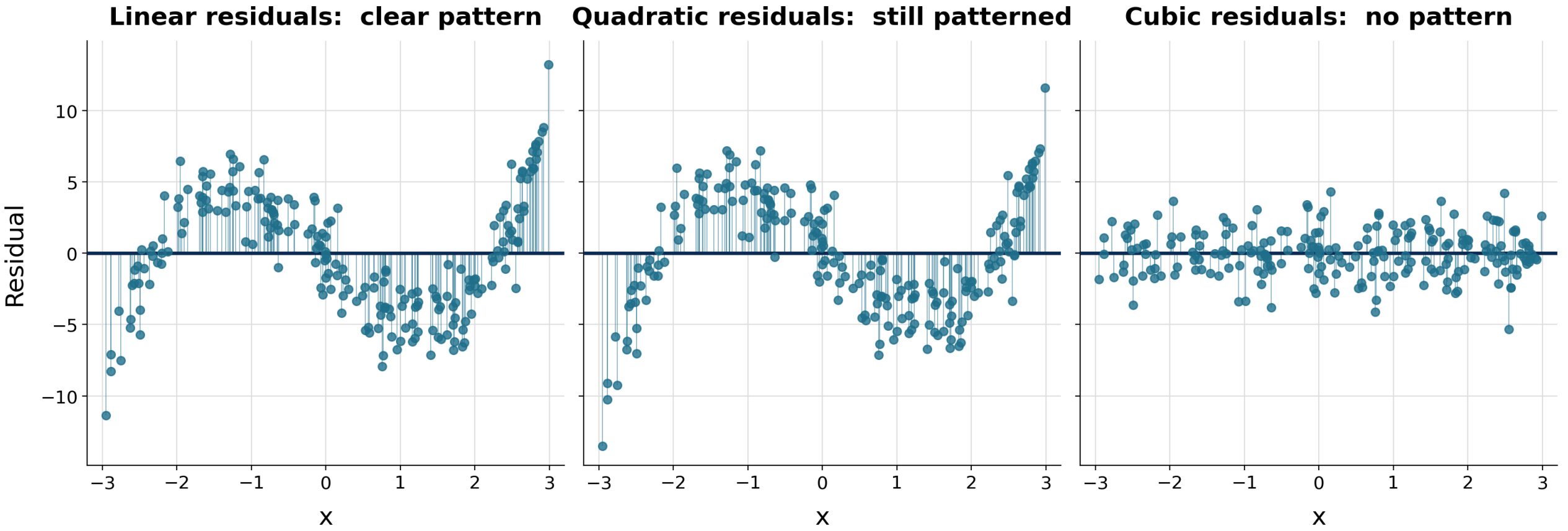




Use **residual plots** for each variable, and **check for a pattern**

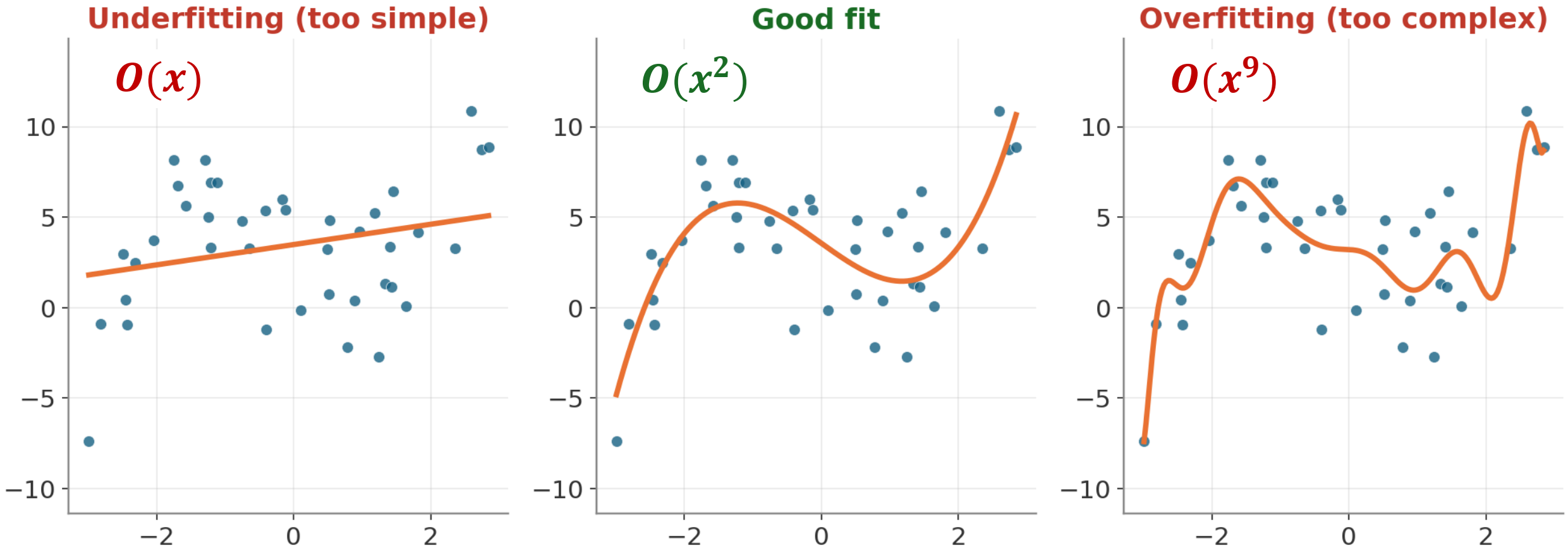
Generalizing to higher-order terms

Add higher-order terms until the residuals look random



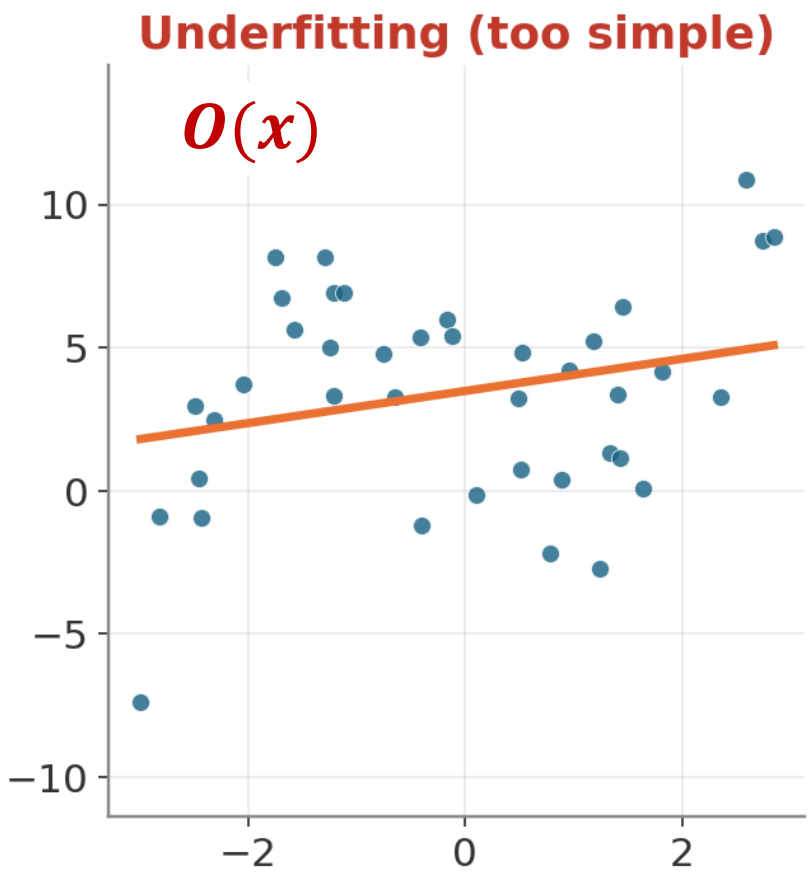
Patterned residuals (a curve) → the model is too simple

Including too many degrees leads to overfitting



Too simple (low-degree) misses the trend; too complex (high-degree) memorizes the noise

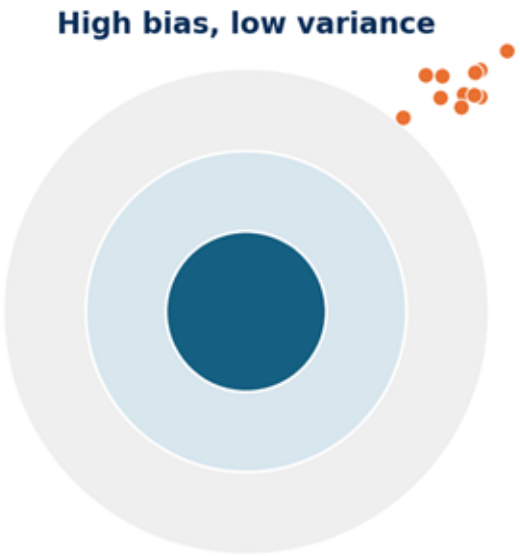
Underfitting *vs.* Overfitting

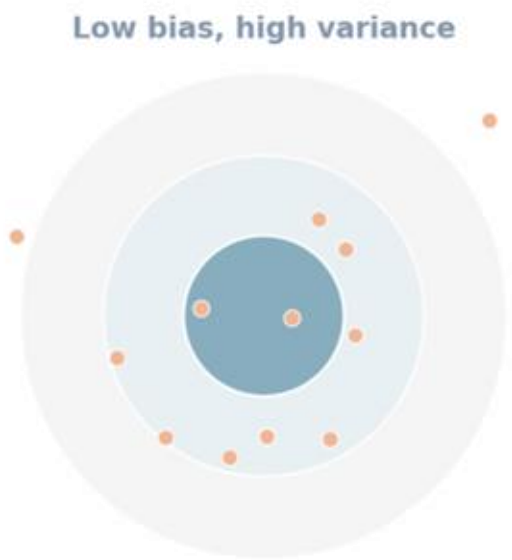




Underfitting:

- Error will be large on the training *and* test data
- **High bias:** we are *systematically* wrong



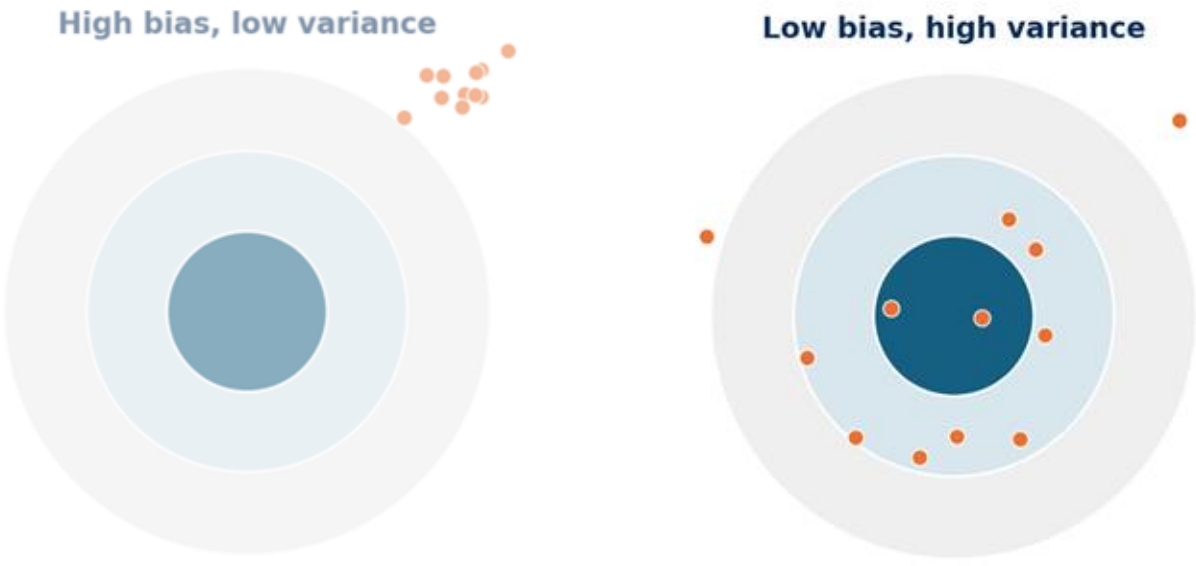


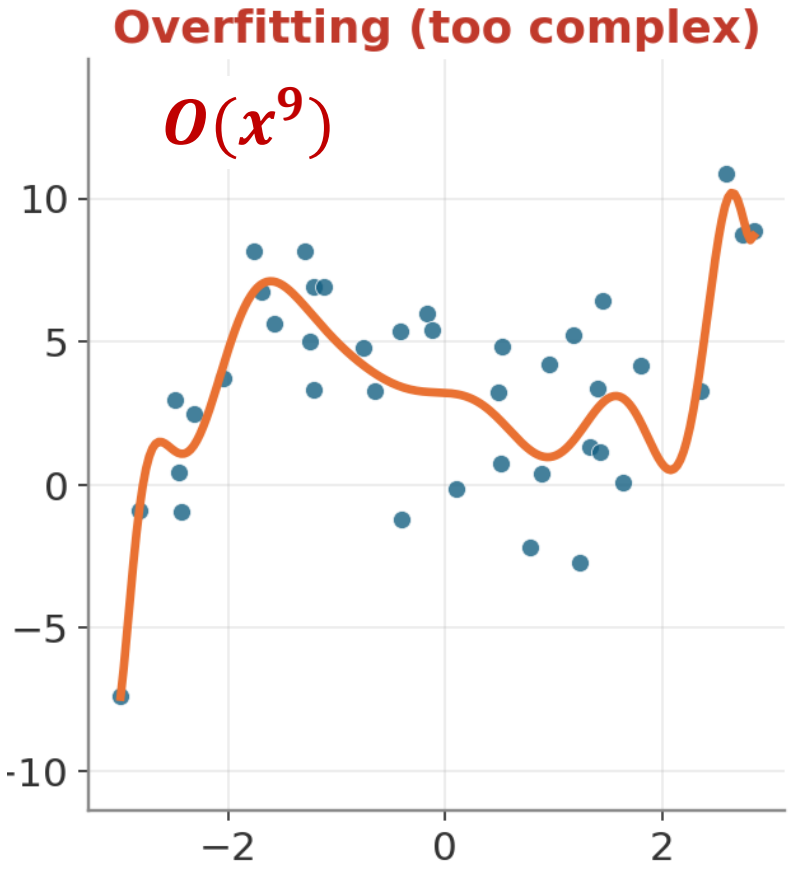
Underfitting *vs.* Overfitting



Overfitting:

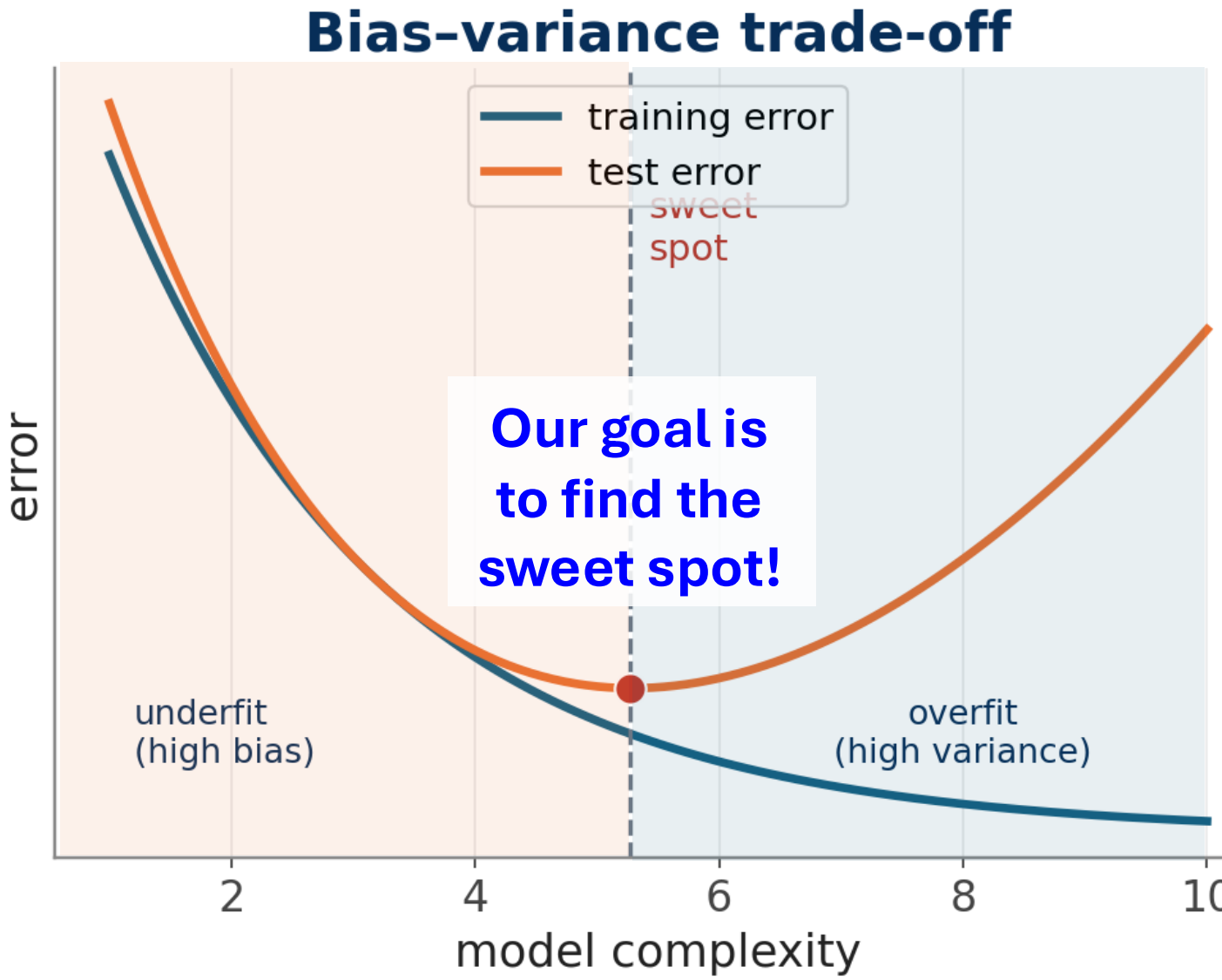
- Error will be **low on training** but **high on test data**
- **High variance:** predictions will change on different subsets of the data





Underfitting *vs.* Overfitting

Underfit:
$\beta_0 + \beta_1 x$



Our goal is to find the sweet spot!

Next time: we'll use regularization to help us do this

Overfit:

$$\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4 + \beta_5 x^5 + \beta_6 x^6 + \beta_7 x^7 + \beta_8 x^8$$

Questions?

What to know for HW and midterm

- How to **calculate β_i using matrix math (normal equation)**
- How to ***interpret* β_i as the slope in x_i holding other features fixed**
- Understand **interaction terms ($\beta_i x_1 x_2$ …)**
- What is **collinearity**, and why should we avoid it
- How many terms to include – effects on **model complexity and overfitting**