Fitting equations to data

SDS 323
James Scott (UT-Austin)

Reference: "Data Science" Chapter 2.

Equations from data

So far we've concentrated on relatively simple visual and numerical summaries of data sets.

In many cases we will want to go further, by fitting an explicit equation, called a *regression model* that describes how one variable (y) changes as a function of some other variables (x).

This is process is called *regression* or *curve fitting* or *supervised learning*: estimating a best guess for y, given x—a *conditional expected value*, $E(y \mid x)$.

Equations from data

For example, you may have heard the following rule of thumb: to calculate your maximum heart rate, subtract your age from 220.

We can express this rule as an equation:

$$\text{MHR} = 220 - \text{Age}$$

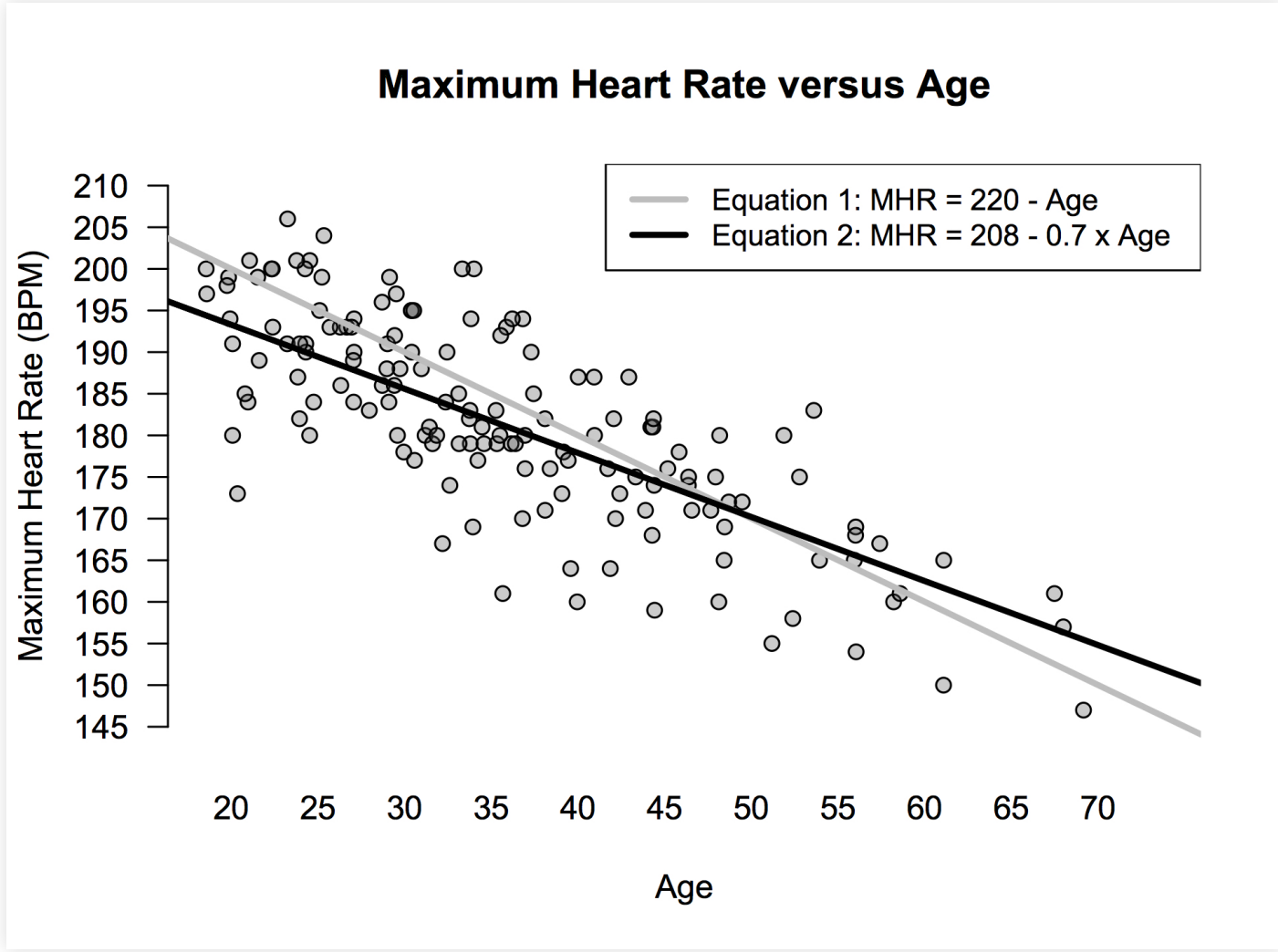
Equations from data

This equation comes from *data*. The study probably went something like this:

- recruit a bunch of people of varying ages
- give them heart rate monitors
- tell them to run really fast on a treadmill
- record their maximum heart rate

Data from this kind of study looks like this…

Equations from data



It turns out that Equation 2 (MHR = 208 - 0.7 × Age) is a better equation: it makes smaller errors, on average.

Equations from data

Top three reasons for fitting an equation:

- to make a prediction
- to summarize the trend in a data set
- to make fair comparisons that adjust for the systematic effect of some important variable. (This is kind of like handicapping in golf.)

Our heart rate example illustrates all three of these concepts.

Making a prediction

Alice is 28. What is her predicted max heart rate?

Our equation expresses the *conditional expected value* of MHR, given a known value of age:

$$\mathrm{E}(\mathrm{MHR} \mid \mathrm{Age}) = 208 - 0.7 \cdot 28 = 188.4$$

This is our best guess without actually putting Alice on a treadmill test until she vomits.

Summarizing a trend

How does max heart rate tend to change with age?

$$\mathrm{E}(\mathrm{MHR} \mid \mathrm{Age}) = 208 - 0.7 \cdot \mathrm{Age}$$

So about 0.7 BPM slower, on average, with every additional year we age.

This isn't a guarantee that *your* MHR will decline at this rate; it's just a population-level average.

Making fair comparisons

A third, very common use of regression modeling is to make *fair comparisons* that adjust for the systematic effect of some common variable.

It's kind of like asking: how big of a head start should The Freeze get?

Making fair comparisons

After the Freeze's Opening Day victory, Usain Bolt declared he's 'too quick for me'

Making fair comparisons

This is not a fair race!

- The Freeze is a former college track star who missed the U.S. Olympic team by 0.02 seconds.
- The other guy is a random fan in a Braves t-shirt.

To the make the race "fair" (and thus interesting), we need to adjust for how fast we *expect* these guys to run, given what we know about them.

Key fact: regression models are great at estimating conditional expectations.

Making fair comparisons

Let's compare two people whose max heart rates are measured using an actual treadmill test:

- Alice is 28 with a maximum heart rate of 185.
- Abigail is 55 with a maximum heart rate of 174.

Clearly Alice has a higher MHR, but let's make things fair! We need to give Abigail a "head start," since max heart rate declines with age.

So who has a higher maximum heart rate *for her age*?

Making fair comparisons

Key idea: compare actual MHR with expected MHR.

Alice's actual MHR is 185, versus an expected MHR of 188.4

$$\begin{aligned}\text{Actual} - \text{Predicted} &= 185 - (208 - 0.7 \cdot 28) \\ &= 185 - 188.4 \\ &= -3.4\end{aligned}$$

Making fair comparisons

Key idea: compare actual MHR with expected MHR.

Abigail's actual MHR is 174, versus an expected MHR of 169.5

$$\begin{aligned}\text{Actual} - \text{Predicted} &= 174 - (208 - 0.7 \cdot 55) \\ &= 174 - 169.5 \\ &= 4.5\end{aligned}$$

Making fair comparisons

So Abigail has a lower absolute MHR, but a higher *age-adjusted* MHR. Her "head start" was the difference between her and Alice's expected MHRs: 188.4 - 169.5 = 18.9.

The equation that relates MHR to age shows us how to place everyone on a level playing field, regardless of age. There are a lot of synonyms for this idea:

- adjusting for x
- statistically controlling for x
- holding x constant

It's all just subtraction! **Compare the difference between actual and expected outcomes.**

Fitting straight lines

The workhorse here is a *linear model*:

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

Notation:

- i indexes each data point
- y_i is the response and x_i is the predictor (feature) for data point i
- β_0 and β_1 are the *parameters* of the model.
- e_i is the *model error* or *residual*, where $E(e_i) = 0$.
- The conditional expected value of y is

$$\hat{y} = E(y \mid x) = \beta_0 + \beta_1 x$$

Fitting straight lines

"Fitting a model" = choosing β_0 and β_1 to make the model errors as small as possible on your data set. In practice "as small as possible" means "least squares." Define the loss function

$$l(\beta_0, \beta_1) = \sum_{i=1}^{N} e_i^2$$

$$= \sum_{i=1}^{N} [y_i - (\beta_0 + \beta_1 x_i)]^2$$

Ordinary least squares (OLS): choose β_0 and β_1 to make $l(\beta_0, \beta_1)$ as small as possible, i.e. to minimize the sum of squared model errors.

Fitting straight lines

It's a straightforward calculus problem to show that the OLS solution is:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}\,,$$

Don't bother memorizing this! Every statistical package on the planet has it built in.

Fitting straight lines

Instead of formula gazing, let's focus on using the fitted equation for our three goals:

- making a prediction
- summarizing the trend in the data
- statistical adjustment: making fair comparisons that adjust for some systematic effect

Example: Austin food critics

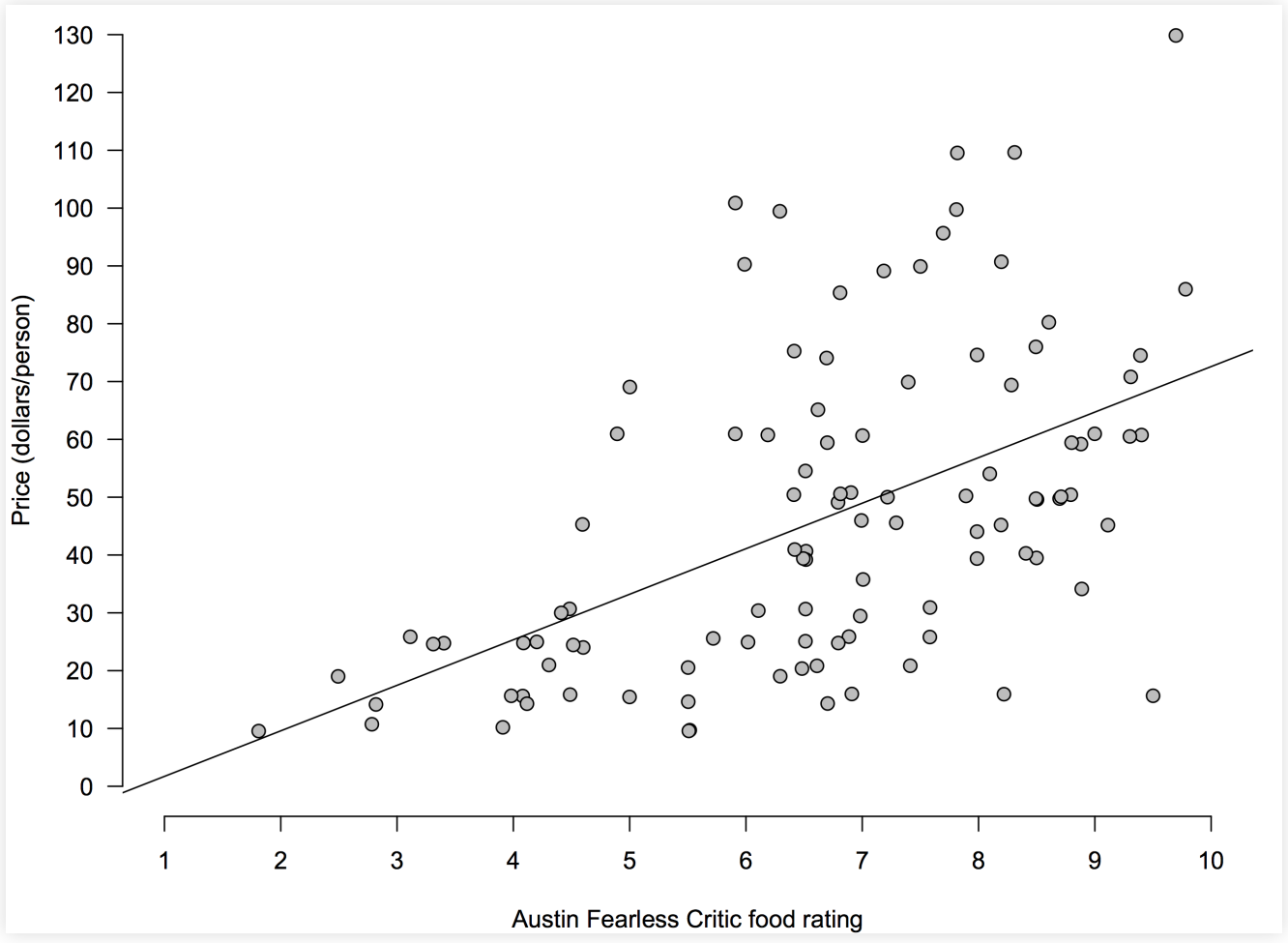
Here are a few lines from a data set on Austin restaurants c.2013:

Name	Type	FoodScore	FeelScore	Price
Franklin Barbecue	Barbecue	9.5	5.5	15
Kerbey Lane Cafe	Vegefusion	6.5	8.5	20
Shoal Creek Saloon	Southern	4.6	8.5	25
Uchi	Japanese,Modern	9.8	8.5	85
Second Bar + Kitchen	Modern	8.7	9.0	50
Lamberts	Southwestern	8.5	9.0	75

- Price = average price of dinner and drinks for one
- FoodScore = critics' rating out of 10

Example: Austin food critics

Our fitted line (from OLS):



$$\text{Price} = -6.2 + 7.9 \cdot \text{FoodScore} + \text{Error}$$

Making a prediction

Suppose you're opening a new restaurant and you've hired a chef with a proven track record of cooking at a 7.5 level.

What price would you expect the Austin market to support for an average meal at your restaurant?

Making a prediction: the algebra

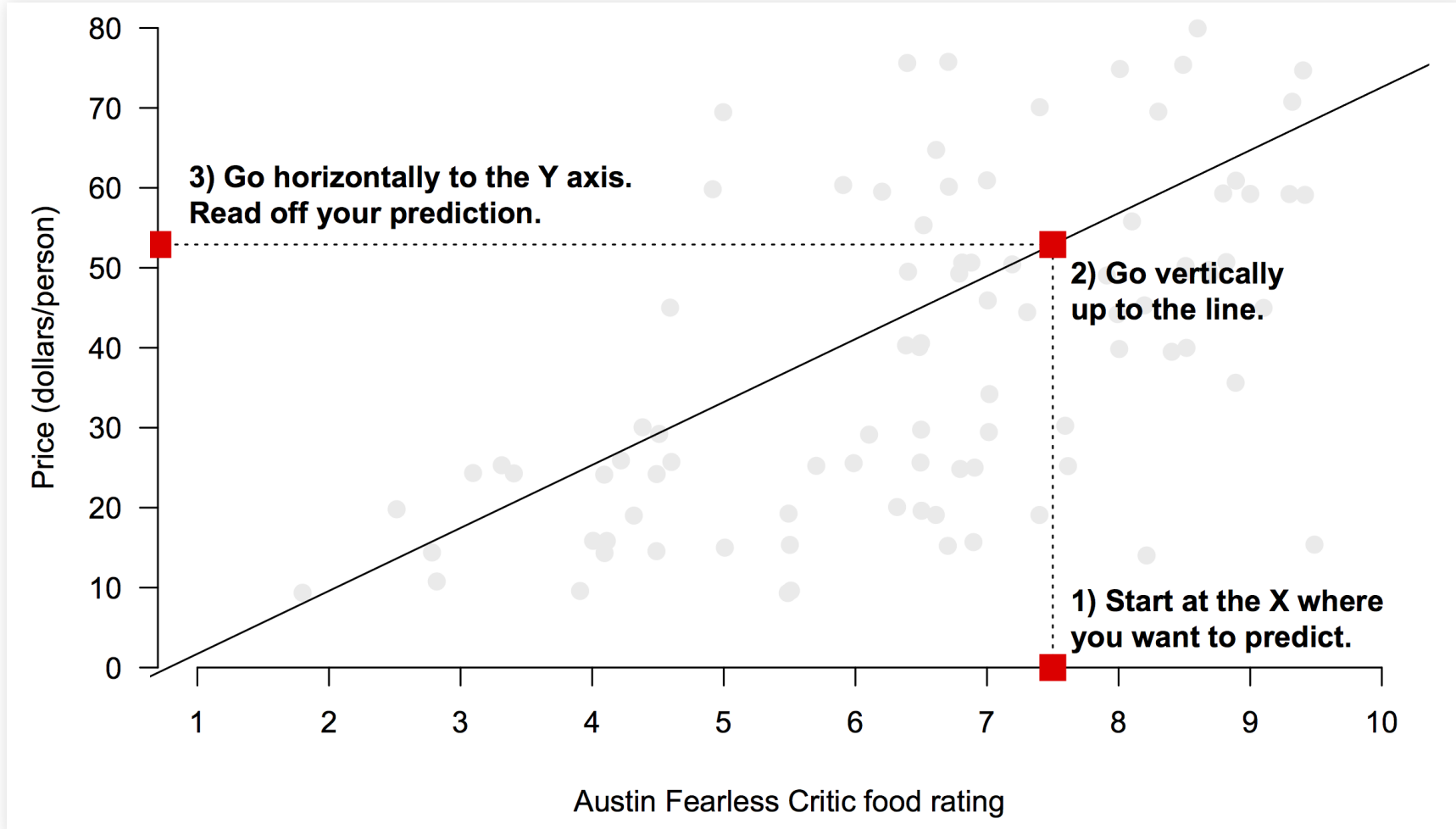
We know that $x = 7.5$. Remember that our fitted equation expresses the *conditional expected value* of y, for a known value of x.

So in light of our data on the Austin market, our best guess for price is

$$E(y \mid x = 7.5) = -6.2 + 7.9 \cdot 7.5 = 53.05$$

Maybe a sensible starting point for thinking about pricing.

Making a prediction: the geometry



Summarizing a trend

What dollar value does the Austin restaurant market seem to place on one extra point of food deliciousness?

Recall our equation

$$E(y \mid x) = -6.2 + 7.9 \cdot x$$

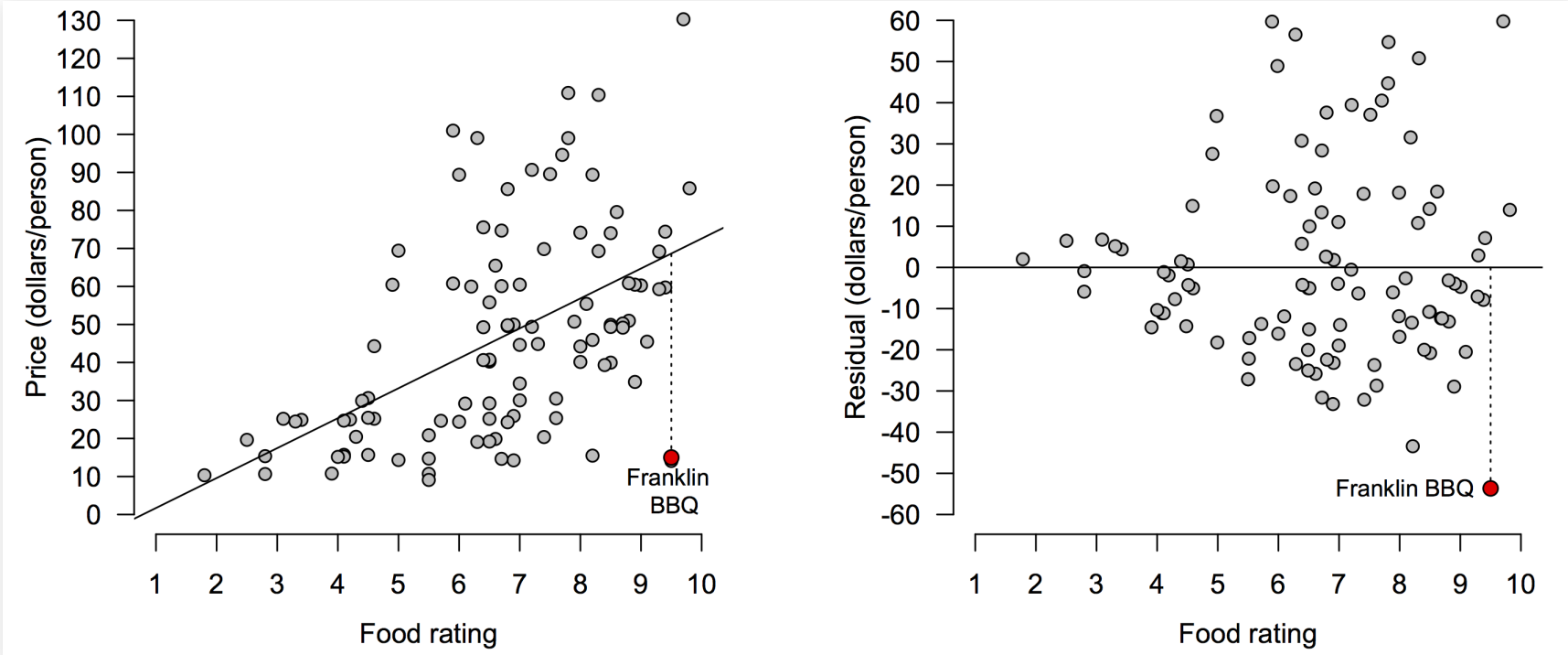
That's $7.90 per point of deliciousness, on average.

Statistical adjustment

What's the "best value" restaurant in Austin, i.e. the one that offers the most delicious food *for its price*?

Recall our key idea: compare actual with predicted.

Statistical adjustment



No surprises here: it's Franklin Barbecue! Actual price is $15 per person; predicted price is nearly $70.

To the code!

Let's dig in to `afc_intro.R` and `afc.csv` on the class website.

Your turn

Download the data in `creatinine.csv` from the course website. Each row is a patient in a doctor's office.

- age: patient's age in years.
- creatclear: patient's creatine clearance rate in mL/minute, a measure of kidney health (higher is better).

Load this data into RStudio and start with a blank script.

Your turn

Use this data, coupled with your knowledge of linear modeling, to answer three questions:

1. What creatinine clearance rate should we expect, on average, for a 55-year-old?
2. How does creatinine clearance rate change with age?
3. Whose creatinine clearance rate is healthier (higher) for their age: a 40-year-old with a rate of 135, or a 60-year-old with a rate of 112?

Beyond straight lines

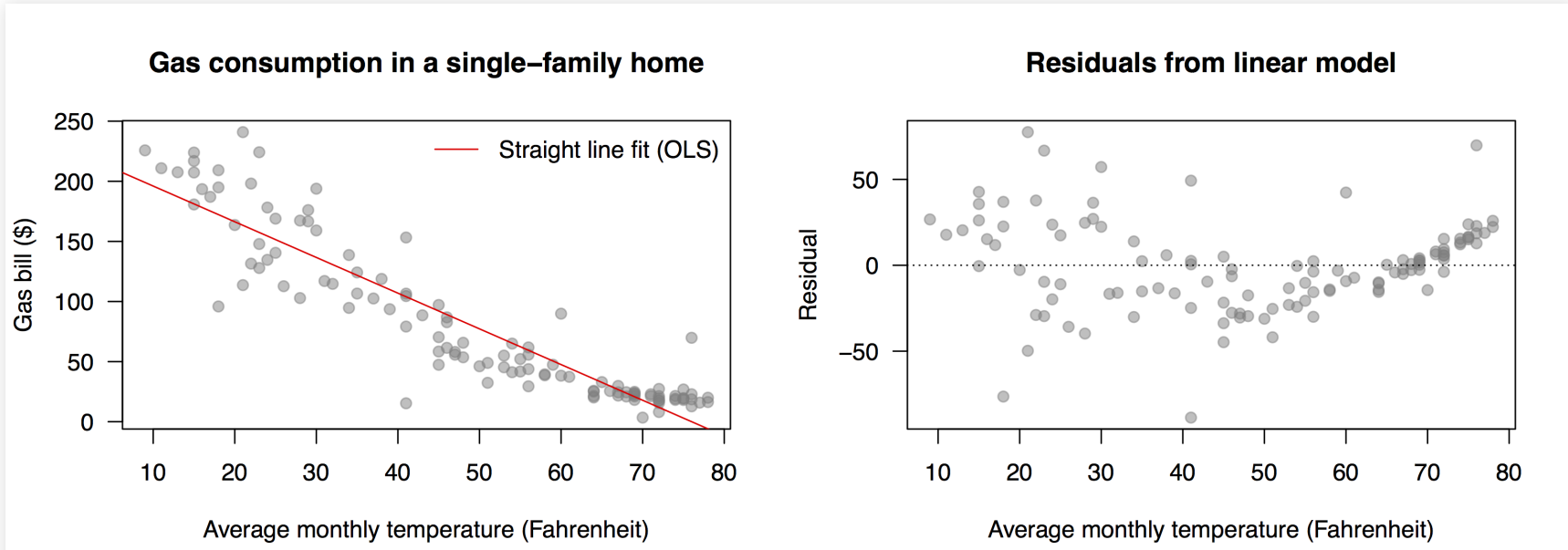
Ordinary least squares is built for linear models. However, we can also use OLS to fit certain kinds of *nonlinear* models. We'll focus on four:

- polynomial models
- piecewise polynomial models (*splines*)
- exponential growth and decay
- power laws

These models are special! (Most nonlinear models cannot be fit using *ordinary* least squares.)

Polynomial models

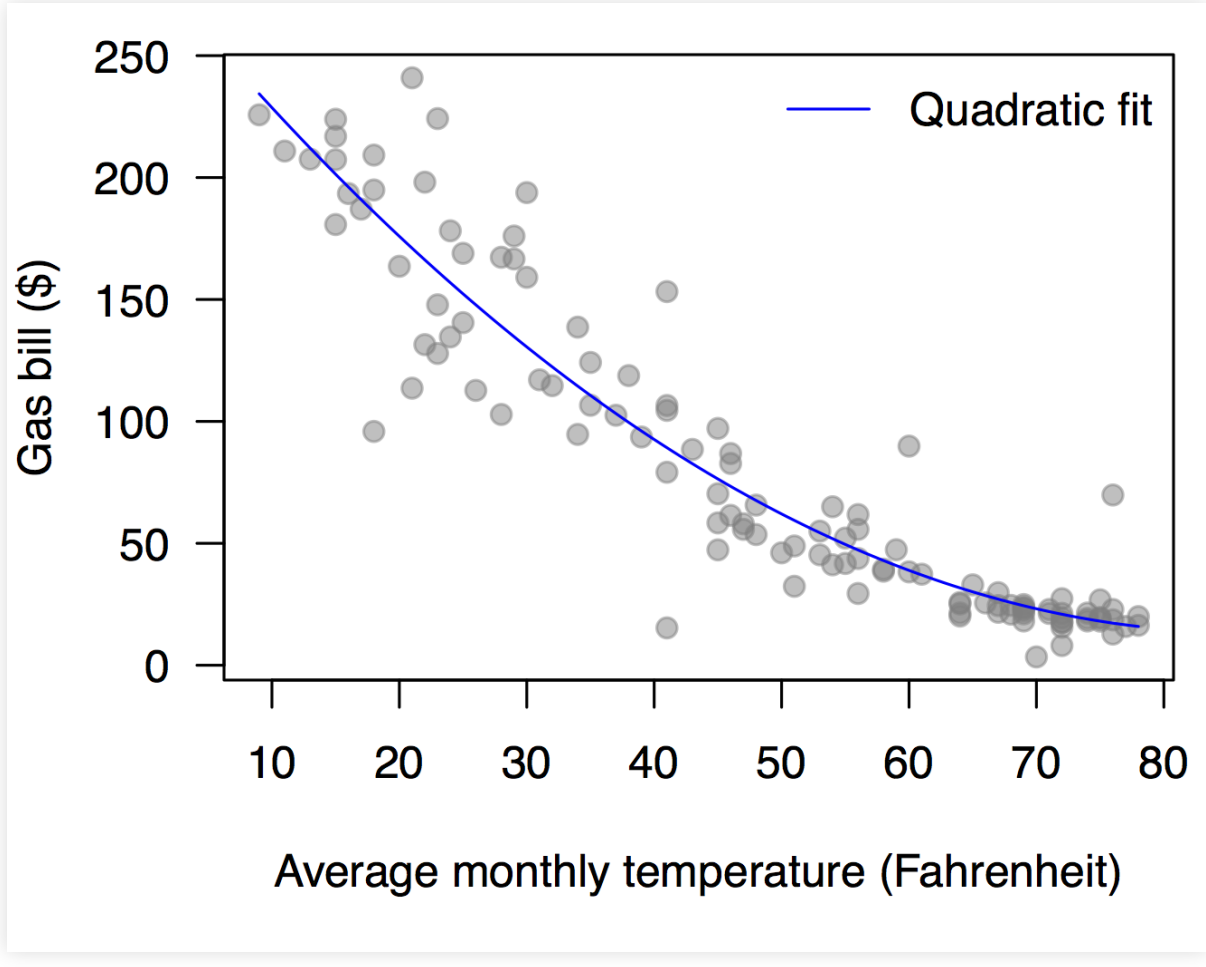
Data on gas consumption versus temperature for a single-family house in Minnesota:



The linear model doesn't fit so well!

Polynomial models

But a quadratic model does!



$$\text{Gas Bill} = \$289 - 6.4 \cdot \text{Temp} + 0.03 \cdot \text{Temp}^2 + \text{Residual}\,.$$

Polynomial models

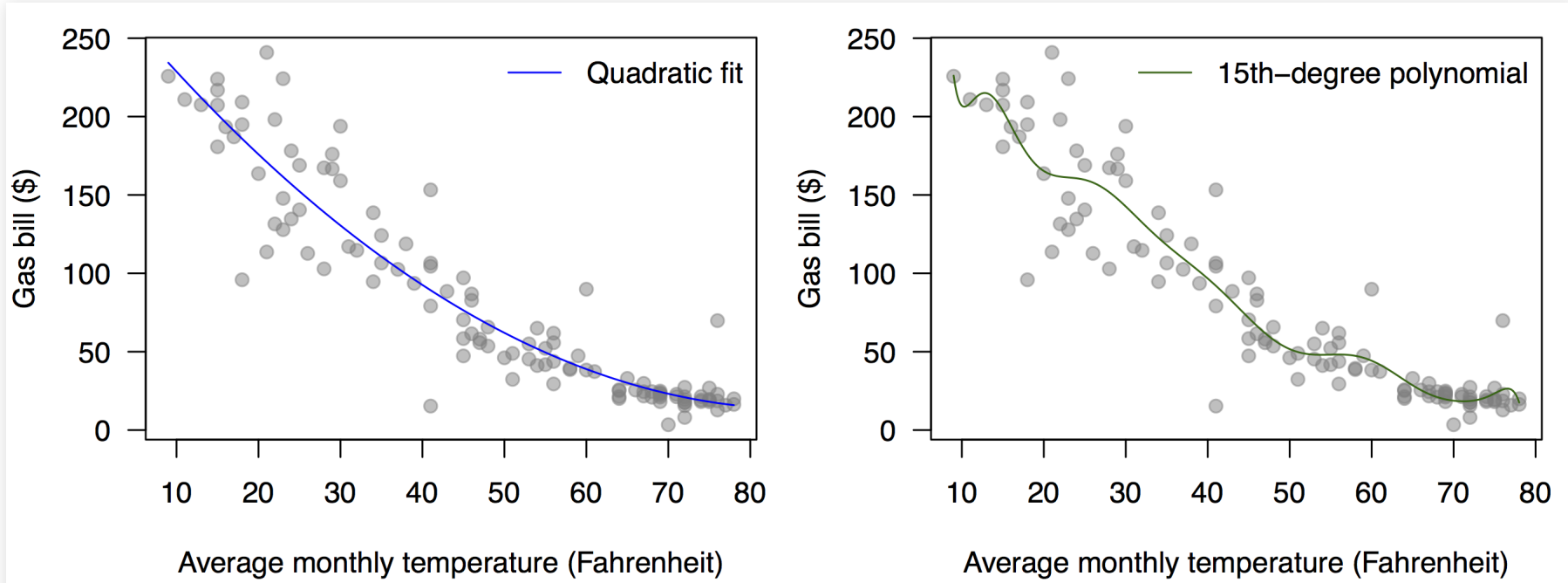
In general, a polynomial model of degree K takes the form

$$E(y \mid x) = \beta_0 + \sum_{j=1}^{K} \beta_j x^j$$

This model is nonlinear in x, but it can still be fit using OLS.

Polynomial models

There is a temptation to get a better fit by choosing a larger K. This can get ridiculous:



Polynomial models: over-fitting

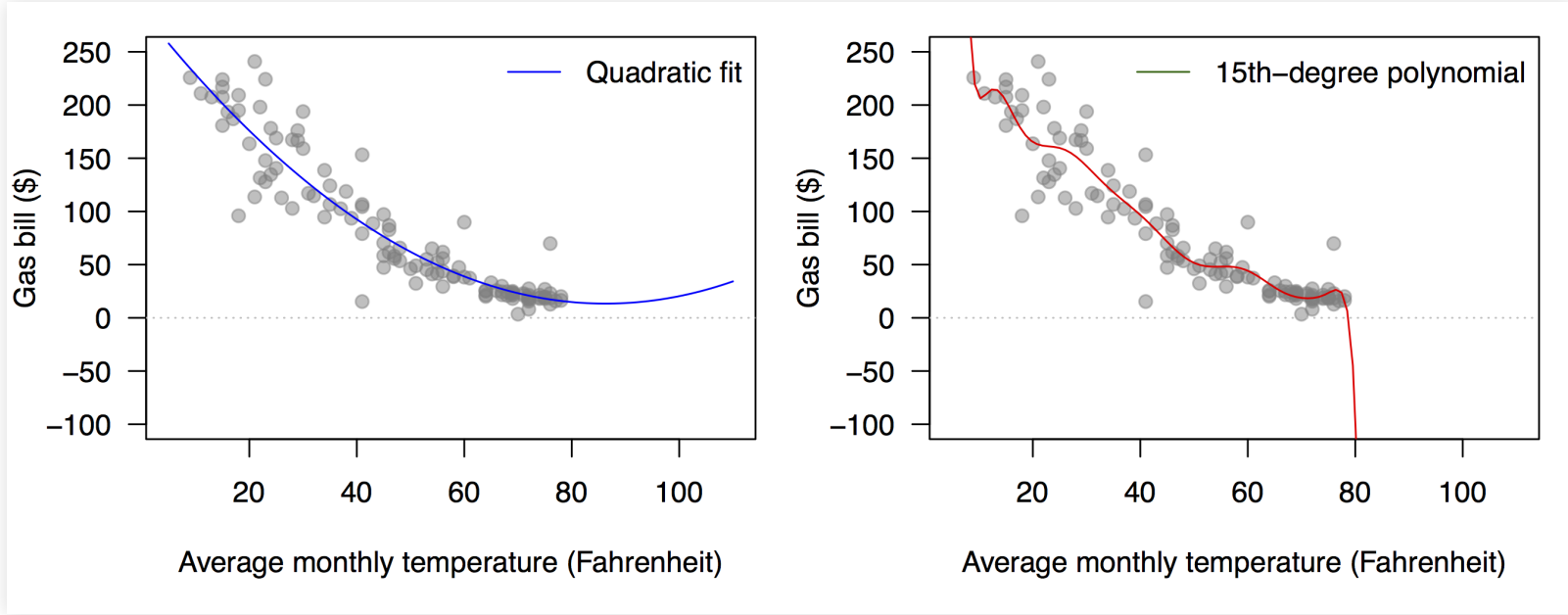
For most data sets, beyond $K = 2$ (quadratic) or $K = 3$ (cubic), we rapidly get into dangerous *over-fitting* territory:

Over-fitting occurs when a model just memorizes random noise in the data set, rather than describes the systematic relationship between x and y.

Severely overfit models usually have one or two dead giveaways:

- non-intuitively wiggly interpolation behavior
- crazy extrapolation behavior!

Polynomial models: over-fitting



In later courses, you'll learn formal diagnostics for over-fitting. In the meantime: you'll pretty much know it when you see it.

Let's dive in to `utilities.csv` and `utilities.R`.

Polynomial models: splines

Another nice extension: piecewise-polynomial models, also known as splines.

To fit a spline, we divide the range of the x variable into nonoverlapping interals $I_0, I_1, I_2, \cdots, I_K$. The breakpoints between the intervals are called *knots*.

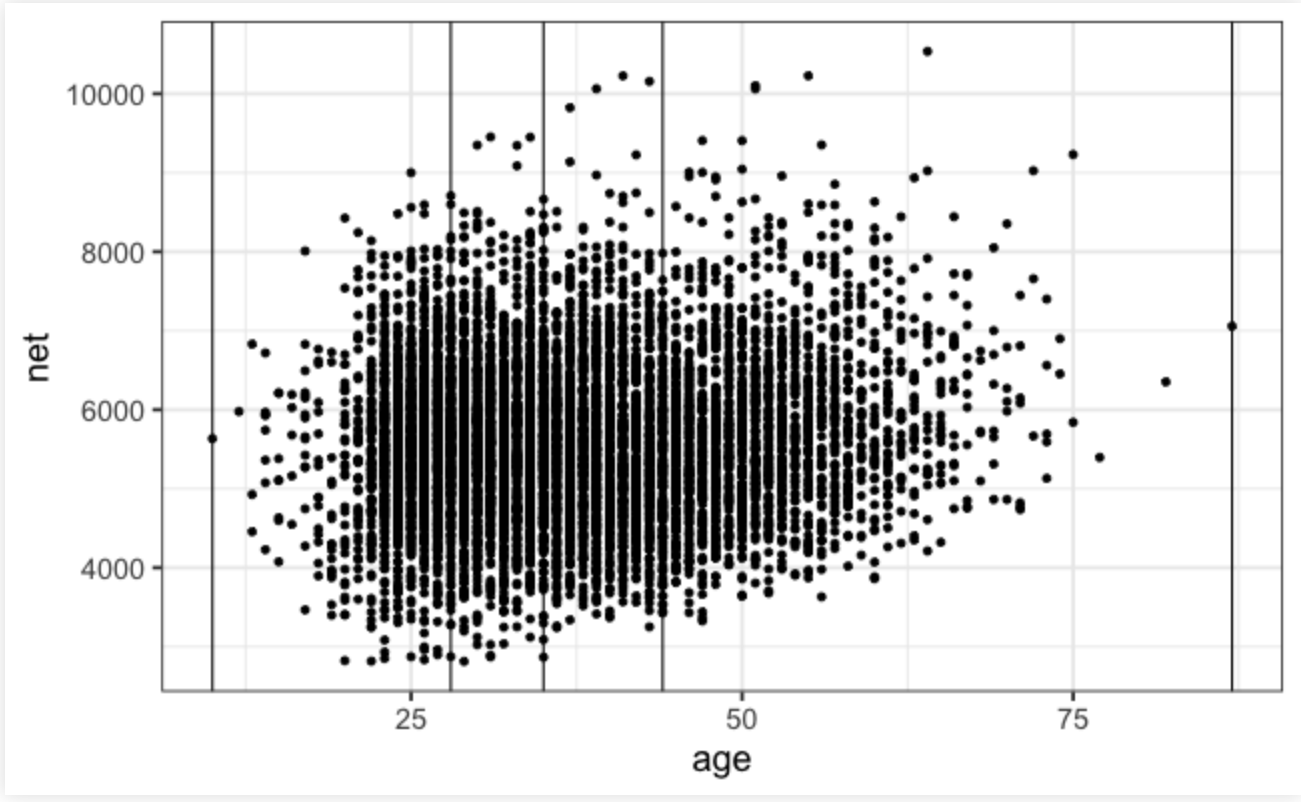
We then fit a polynomial equation separately on each interval, "gluing" the individual polynomials together so that the overall curve is smooth.

Polynomial models: splines

Here's some data on finishing times from runners in the 10-mile Cherry Blossom Road Race in Washington, D.C., held every April:

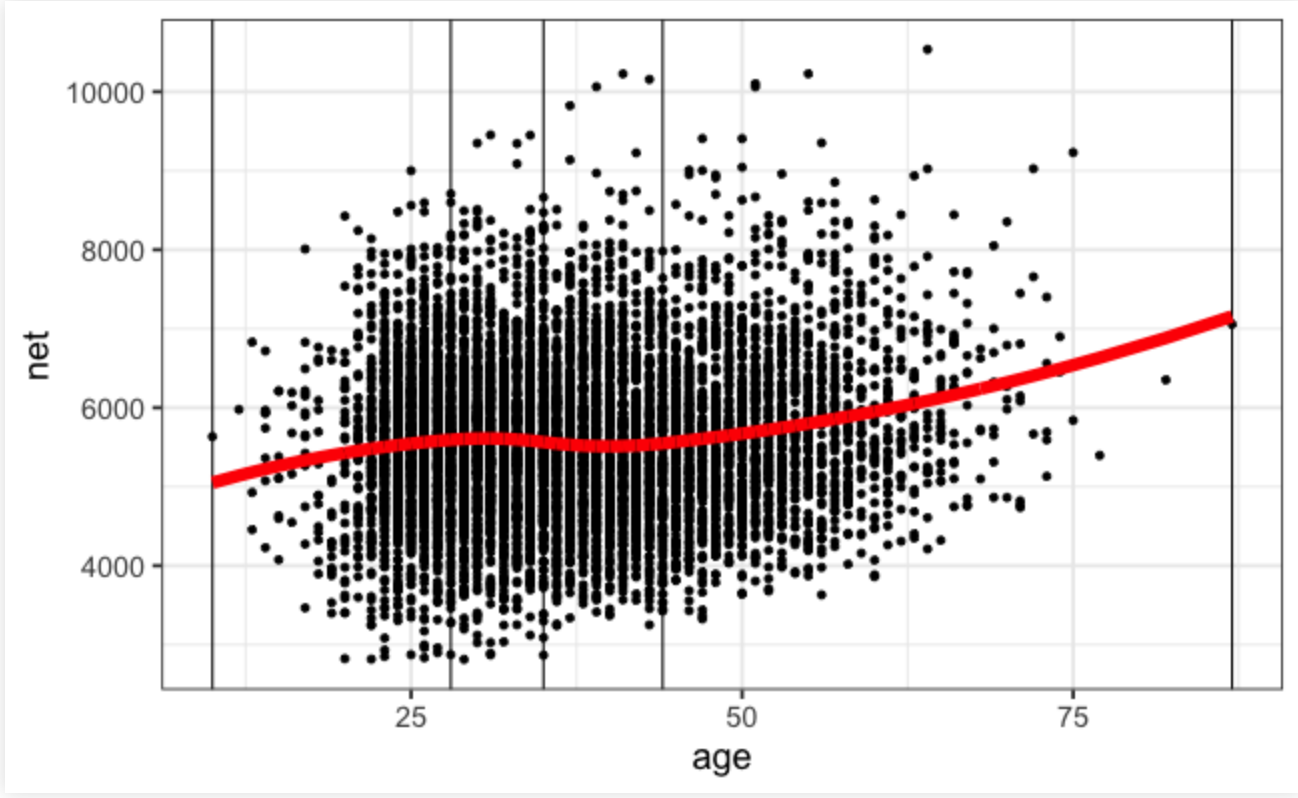
```
  state time  net age sex
1    DC 6070 5669  35   M
2    DC 5086 4898  31   F
3    VA 6135 5895  23   F
4    DC 4886 4673  42   F
5    MD 6440 6225  39   M
6    MD 6185 5748  41   M
7    VA 5118 5018  35   M
8    VA 6481 6034  45   M
```

Polynomial models: splines



Three knots create four disjoint intervals (knots at the 25th, 50th, and 75th percentiles of temperature).

Polynomial models: splines



Separate polynomials on each interval, glued together in a smooth fashion. **What might explain the non-monotone behavior?** (Code in `race_splines.R`.)

Polynomial models: splines

There are probably at least two things going on here:

- actual non-monotonicity: runners reach their physical peak fairly late compared to other athletes (sometime in their 30s or 40s).
- survivorship bias: only the more serious runners are still running in their late 30s and through their 40s.

Bottom line: the flexible piece-wise polynomial (spline) model allowed us to **practice good data science**. We could:

- see these subtle effects in the data.
- form theories about what might be causing them.

Exponential growth and decay

An exponential growth or decay model looks like this:

$$y = \alpha e^{\beta_1 x}$$

where:

- α is a baseline level of the response y at $x = 0$.
- β_1 describes the rate of growth (+) or decay (−) for a one-unit change in x.

Exponential growth and decay

This formula comes from an analogy with continuously compounded interest. Suppose you start with α dollars, invested at rate β_1 and compounded n times annually. Then after t years, your investment is worth

$$y = \alpha\left(1 + \frac{\beta_1}{n}\right)^{nt}$$

If we take the limit as n gets large, we get

$$y = \alpha e^{\beta_1 t}$$

Exponential growth and decay

This is a non-linear model, but it's easy to fit using OLS!

Here's the trick: if $y = \alpha e^{\beta_1 x}$, then

$$\begin{aligned}\log(y) &= \log\left(\alpha e^{\beta_1 x}\right) \\ &= \log \alpha + \beta_1 x\end{aligned}$$

This is a linear function after all: not in y versus x, but in $\log(y)$ versus x.

Exponential growth and decay

This gives us a simple recipe: Fit a linear regression model where the y variable has been log-transformed:

$$\log(y_i) = \beta_0 + \beta_1 x_i + e_i$$

This tells us the parameters of our exponential growth or decay model:

- $\beta_0 = \log \alpha$, so $\alpha = e^{\beta_0}$ is the initial level of y at $x = 0$.
- β_1 is the growth rate per unit change in x.

See example in `ebola.R`.

Power laws

A similar trick works for power laws, where

$$y = \alpha x^{\beta}$$

This is a very common model in microeconomics, where we often use power laws to model change in consumer demand as a function of price:

$$Q = KP^{E}$$

where Q is quantity demanded, P is price, E is price elasticity of demand (PED), and K is a constant.

Exponential growth and decay

We can git a model like this using OLS as well.

Here's the trick: if $y = \alpha x^{\beta_1}$, then

$$\begin{aligned}\log(y) &= \log\left(\alpha x^{\beta_1}\right) \\ &= \log \alpha + \beta_1 \log(x)\end{aligned}$$

This is also linear function, in $\log(y)$ versus $\log(x)$.

Exponential growth and decay

Now we fit a linear regression model with *both* variables log-transformed:

$$\log(y_i) = \beta_0 + \beta_1 \log(x_i) + e_i$$

This tells us the parameters of our exponential growth or decay model:

- $\beta_0 = \log \alpha$, so $\alpha = e^{\beta_0}$ is our leading constant.
- β_1 is the elasticity: if x changes by 1%, then y changes by β_1%.

See example on class website.

Power laws

Your turn! In `milk.csv`, you're given a data set on consumer demand for milk, along with price. Your ultimate goal is to answer the question: how much should the store charge for a carton of milk?

Let's dive in to the case study on milk prices on the class website.

Remember: demand versus price often follows a power law!