GENERAL ASSEMBLY

LINEAR REGRESSION

Mason Gallo, Data Scientist

AGENDA

- Ordinary Linear Regression
- Why we need regularization
- Lasso
- Ridge
- Common terms

OBJECTIVES

- Big picture of Linear Regression
- Learn why we may need to use regularization
- Build your best Linear Regression model

MOTIVATING EXAMPLE: PREDICTING SALES

PREDICTING SALES BASED ON MARKETING MIX

TV	Radio	Newspaper	Sales
230.1	37.8	69.2	22.1
44.5	39.3	45.1	10.4
17.2	45.9	69.3	9.3
151.5	41.3	58.5	18.5
180.8	10.8	58.4	12.9
8.7	48.9	75	7.2
57.5	32.8	23.5	11.8
120.2	19.6	11.6	13.2
8.6	2.1	1	4.8
199.8	2.6	21.2	10.6
66.1	5.8	24.2	8.6

PREDICTING SALES BASED ON MARKETING MIX

TV	Radio	Newspaper	Sales
230.1	37.8	69.2	22.1
44.5	39.3	45.1	10.4
17.2	45.9	69.3	9.3
151.5	41.3	58.5	18.5
180.8	10.8	58.4	12.9
8.7	48.9	75	7.2
57.5	32.8	23.5	11.8
120.2	19.6	11.6	13.2
8.6	2.1	1	4.8
199.8	2.6	21.2	10.6
66.1	5.8	24.2	8.6

WE WILL USE LINEAR REGRESSION TO RECOMMEND A MARKETING MIX

INTRO TO LINEAR REGRESSION

REGRESSION PROBLEMS

	continuous	categorical
supervised	???	???
unsupervised	???	???

	continuous	categorical
supervised	regression	classification
unsupervised	dimension reduction	clustering

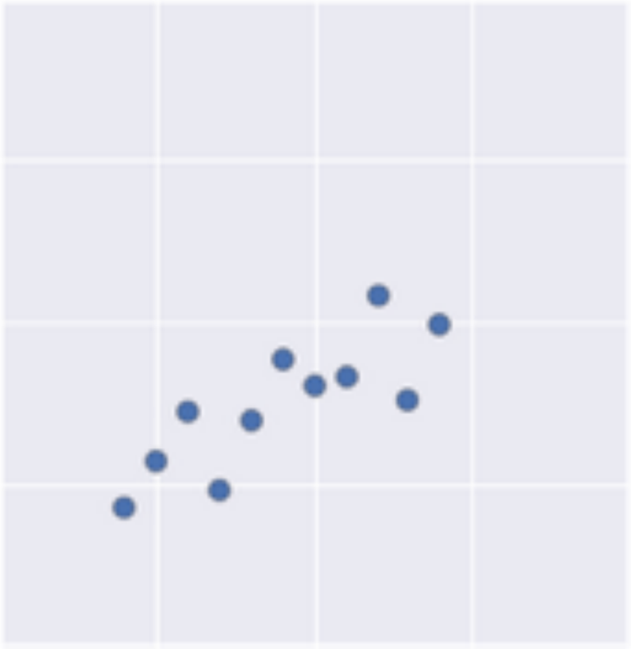
Q: What is a **regression** *model?*

A: A functional relationship between input & response variables

The **simple linear regression** *model captures a linear relationship between a single input variable x and a response variable y:*

Q: What is a **regression** *model?*

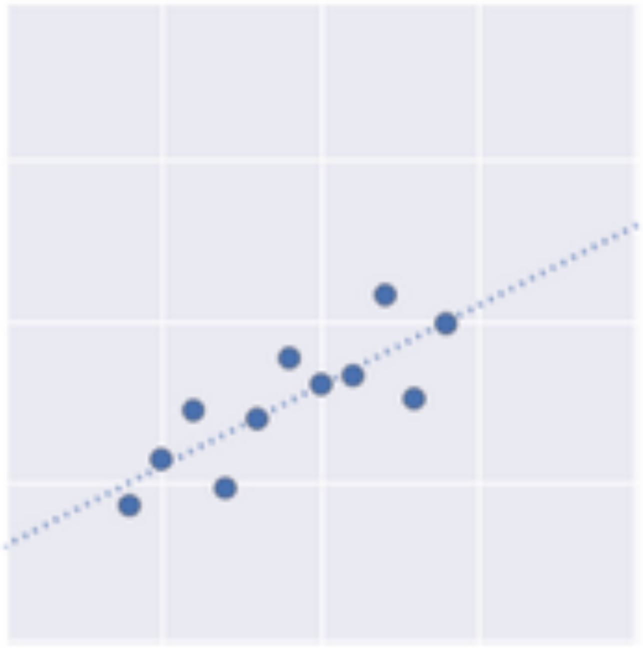
A: A functional relationship between input & response variables



Source: Anscombe's Quartet

Q: What is a **regression** *model?*

A: A functional relationship between input & response variables



Source: Anscombe's Quartet

Q: What is a **regression** *model?*

A: A functional relationship between input & response variables

The **simple linear regression** *model captures a linear relationship between a single input variable* x *and a response variable* y*:*

$$y = \alpha + \beta x + \varepsilon$$

Q: What do the terms in this model mean?

$$y = \alpha + \beta x + \varepsilon$$

Q: What do the terms in this model mean?

$$y = \alpha + \beta x + \varepsilon$$

A: y = **response variable** *(the one we want to predict)*

Q: What do the terms in this model mean?

$$y = \alpha + \beta x + \varepsilon$$

A: y = **response variable** *(the one we want to predict)*

x = **input variable** *(the one we use to train the model)*

Q: What do the terms in this model mean?

$$y = \alpha + \beta x + \varepsilon$$

A: y = **response variable** *(the one we want to predict)*

x = **input variable** *(the one we use to train the model)*

α = **intercept** *(where the line crosses the y-axis)*

Q: What do the terms in this model mean?

$$y = \alpha + \beta x + \varepsilon$$

A: y = **response variable** *(the one we want to predict)*

x = **input variable** *(the one we use to train the model)*

α = **intercept** *(where the line crosses the y-axis)*

β = **regression coefficient** *(the model "parameter")*

Q: What do the terms in this model mean?

$$y = \alpha + \beta x + \varepsilon$$

A: y = **response variable** *(the one we want to predict)*

x = **input variable** *(the one we use to train the model)*

α = **intercept** *(where the line crosses the y-axis)*

β = **regression coefficient** *(the model "parameter")*

ε = **residual** *(the prediction error)*

INTRO TO REGRESSION

$$y = \alpha + \beta x + \varepsilon$$

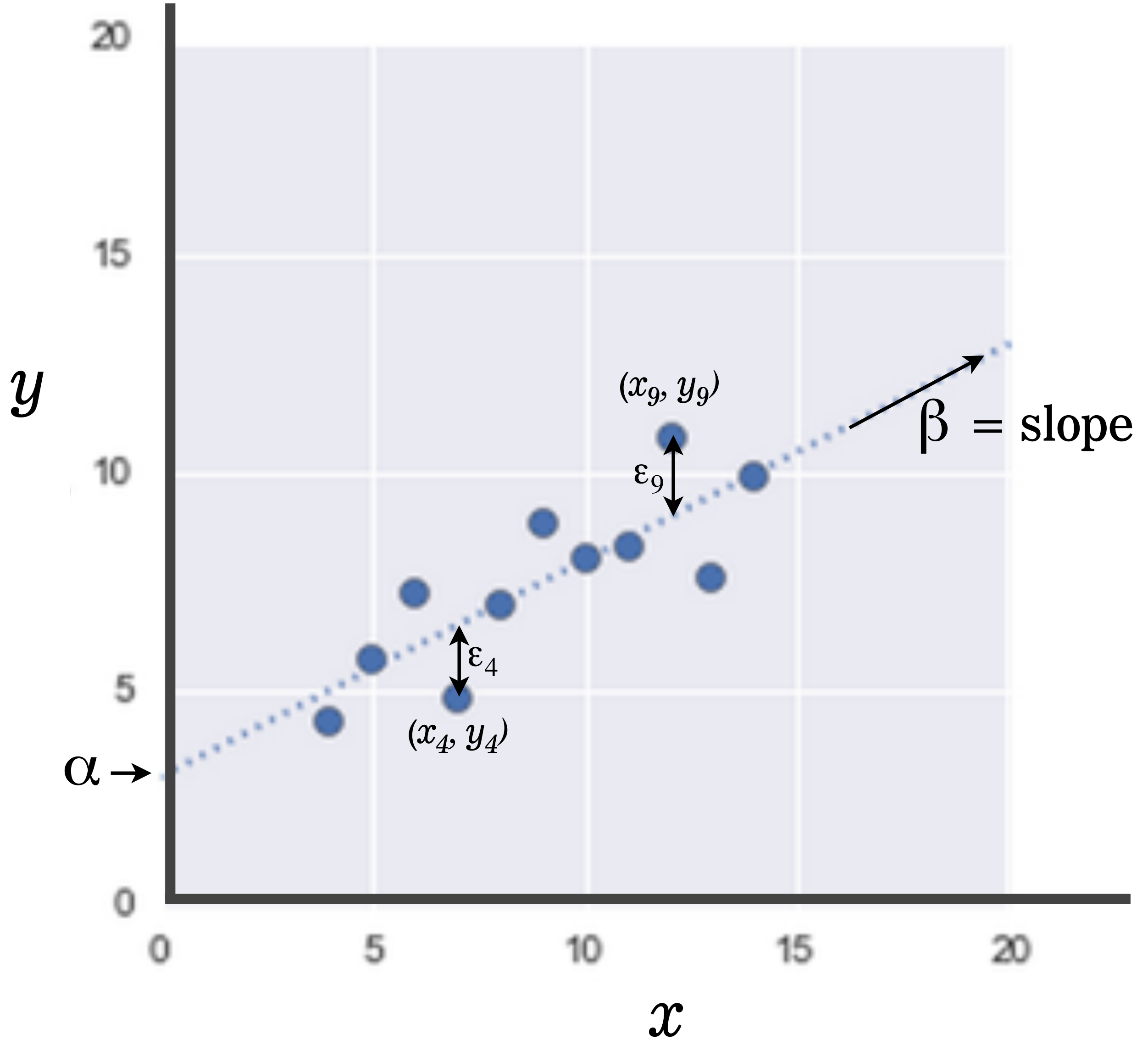
y = **response variable**

x = **input variable**

α = **intercept**

β = **regression coefficient**

ε = **residual** *(the error)*



Source: Anscombe's Quartet

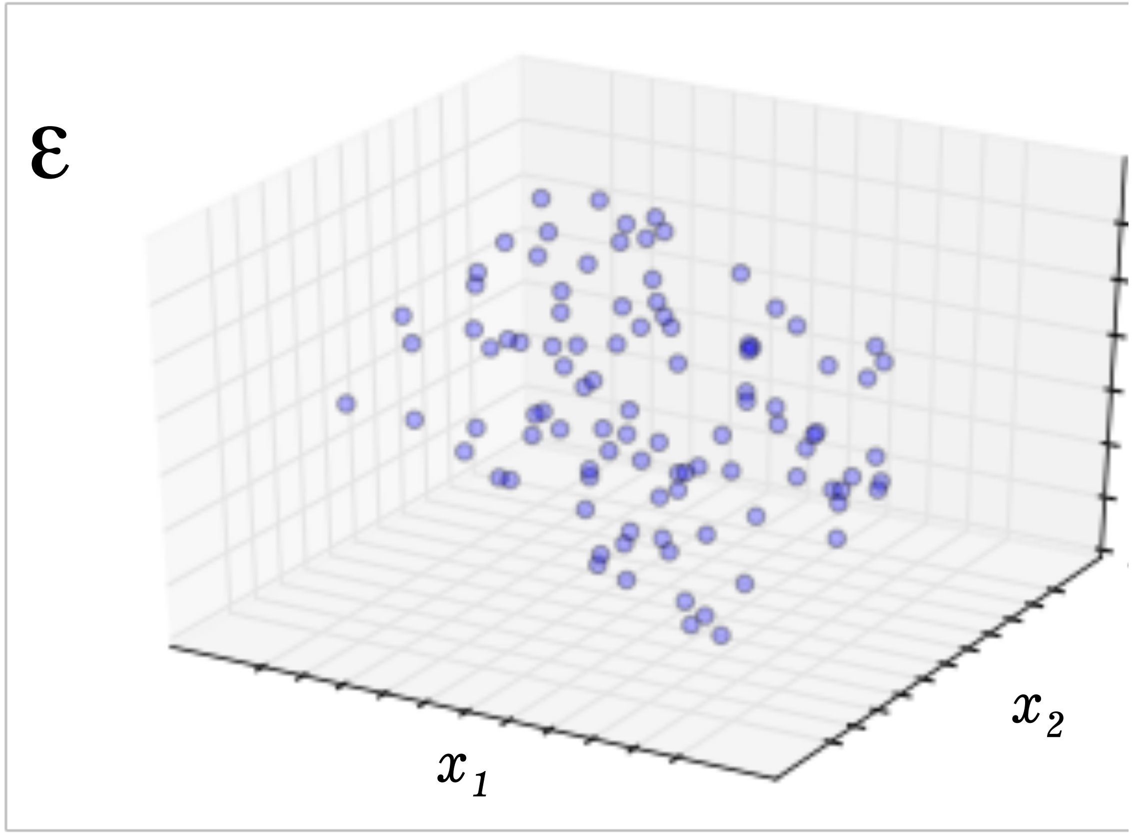
We can extend this model to several input variables, giving us the **multiple linear regression** *model:*

We can extend this model to several input variables, giving us the **multiple linear regression** *model:*

$$y = \alpha + \beta_1 x_1 + \ldots + \beta_n x_n + \varepsilon$$

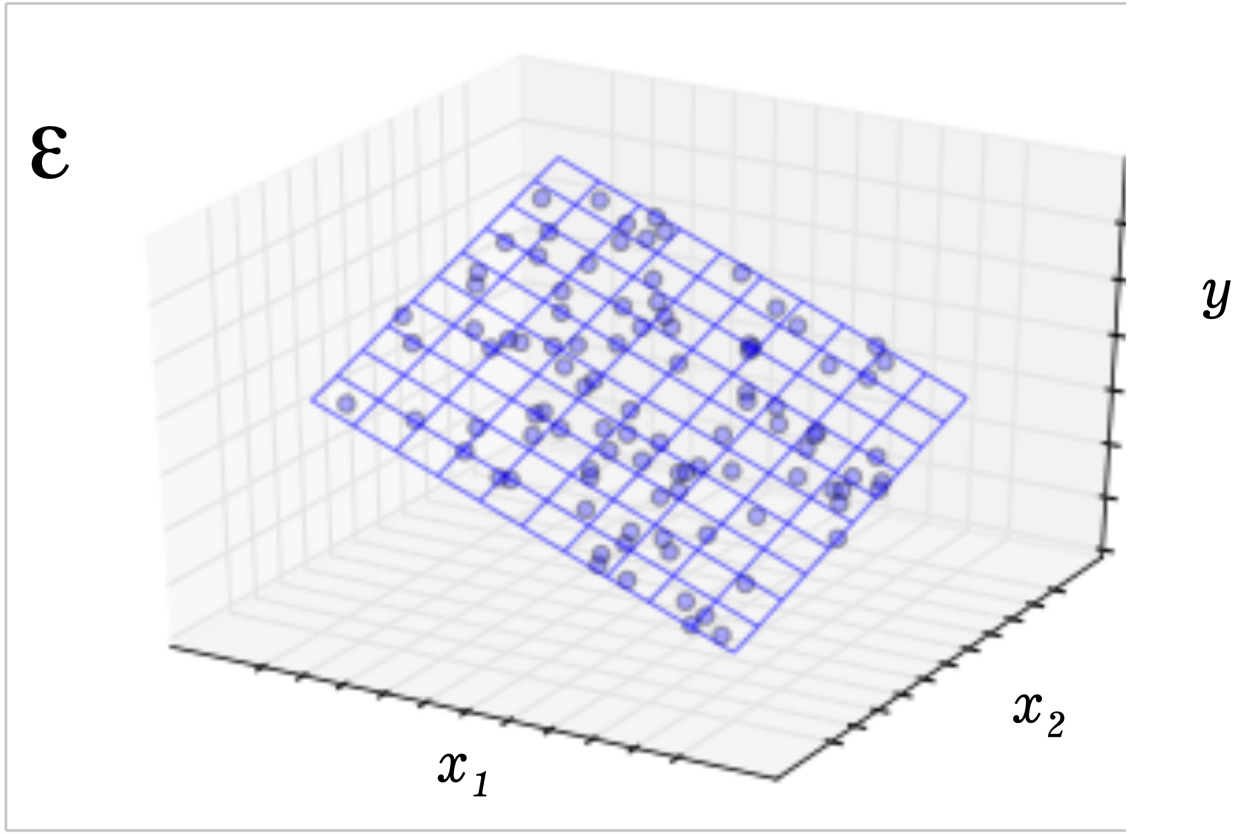
We can extend this model to several input variables, giving us the **multiple linear regression** *model:*

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$



We can extend this model to several input variables, giving us the **multiple linear regression** *model:*

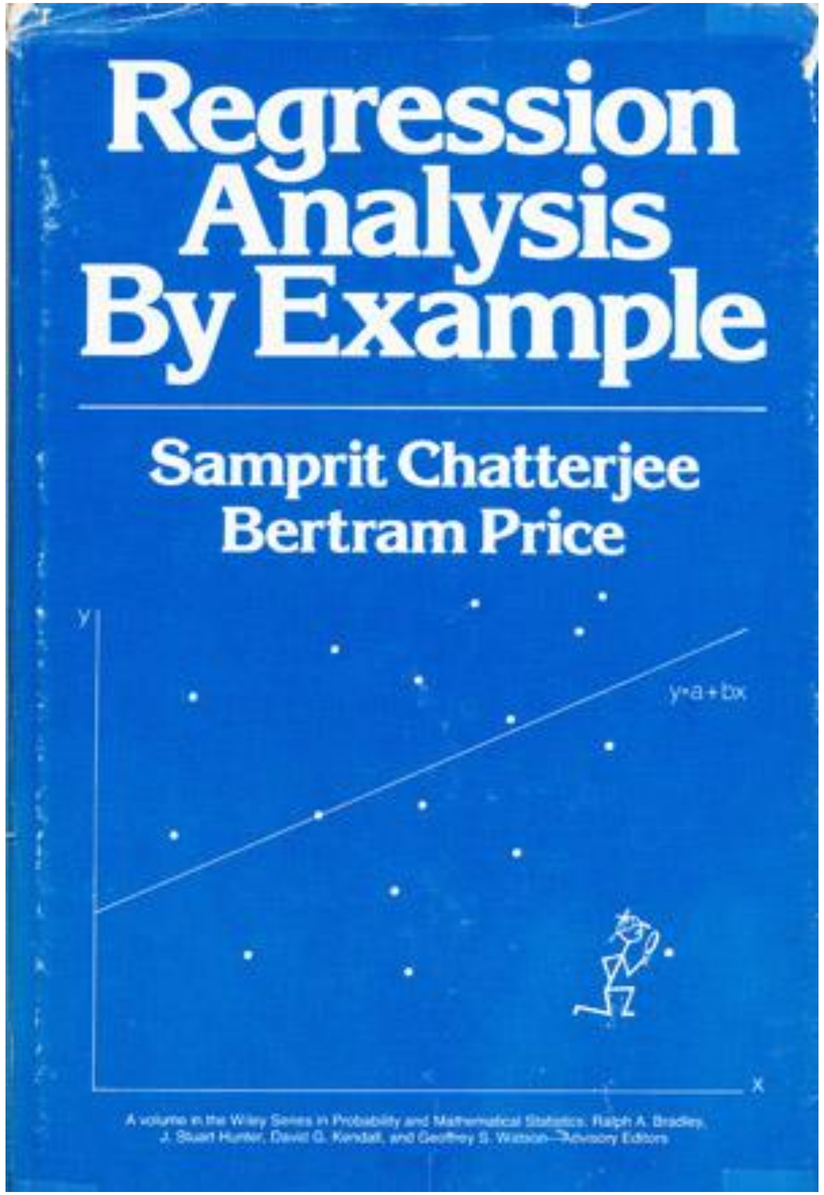
$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

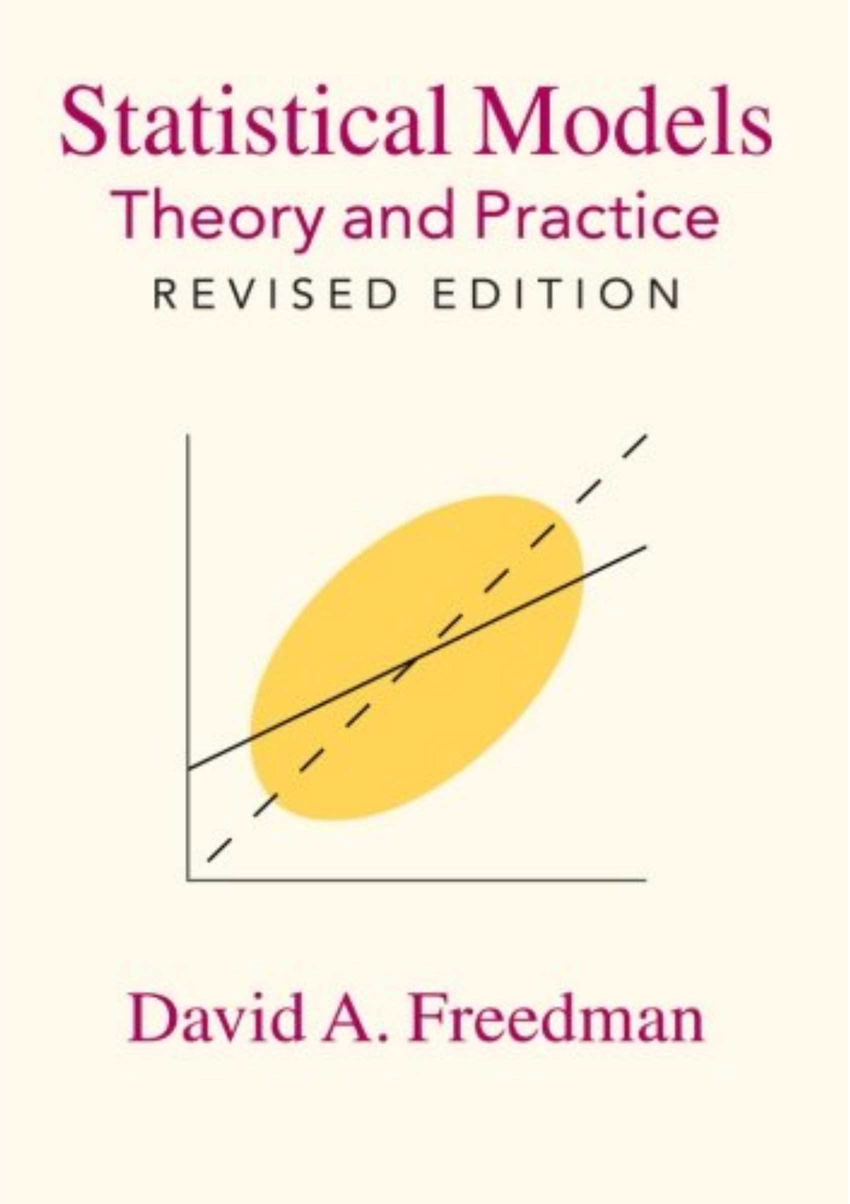


Linear regression involves several technical assumptions and is often presented with lots of mathematical formality.

The math is not very important for our purposes, but you should check it out if you get serious about solving regression problems.

INTRO TO REGRESSION





Q: How do we fit a regression model to a dataset?

Q: How do we fit a regression model to a dataset?

A: In theory, minimize the sum of the squared residuals (OLS).

Q: How do we fit a regression model to a dataset?

A: In theory, minimize the sum of the squared residuals (OLS).

In practice, any respectable piece of software will do this for you.

Q: How do we fit a regression model to a dataset?
A: In theory, minimize the sum of the squared residuals (OLS).

In practice, any respectable piece of software will do this for you.

But again, if you get serious about regression, you should learn how this works!

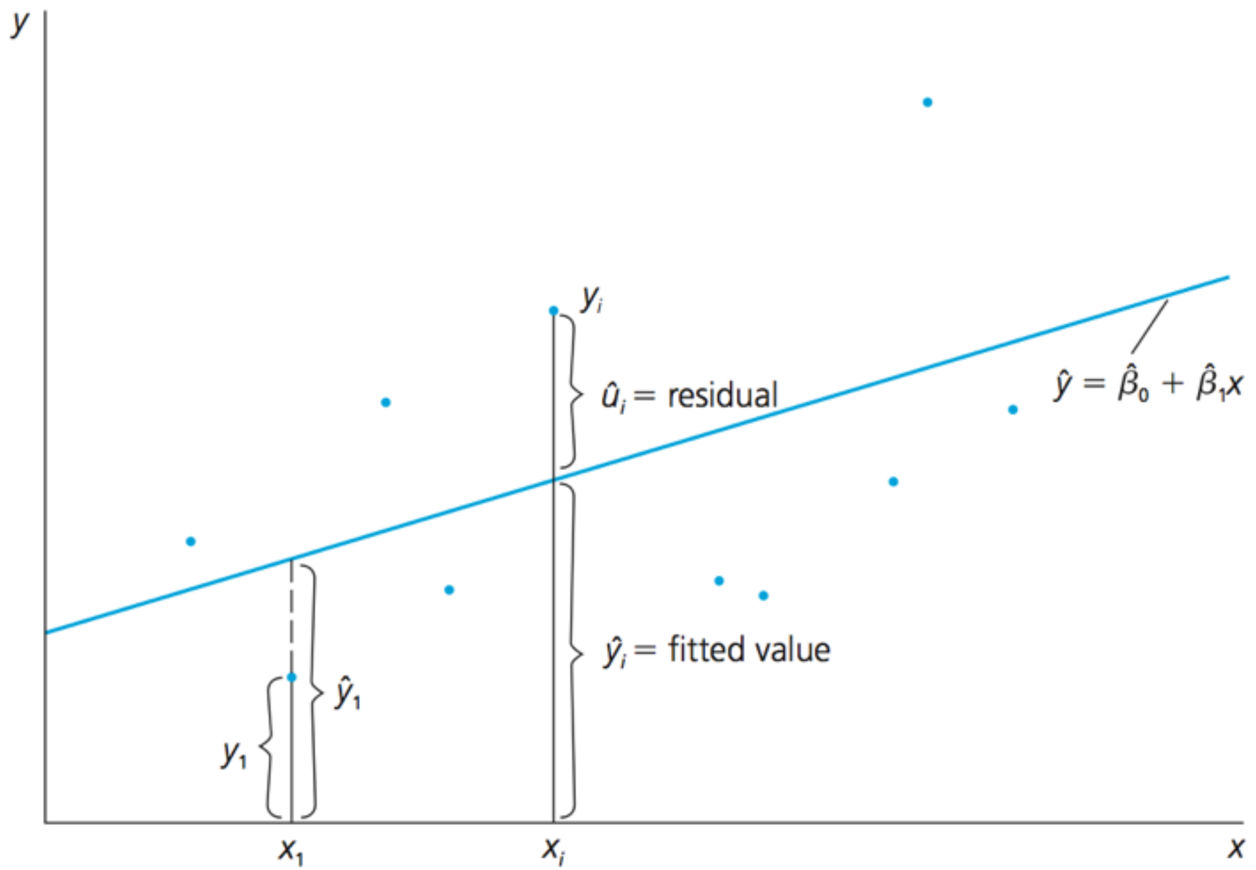
Q: How do we fit a regression model to a dataset?

A: In theory, minimize the sum of the squared residuals (OLS).

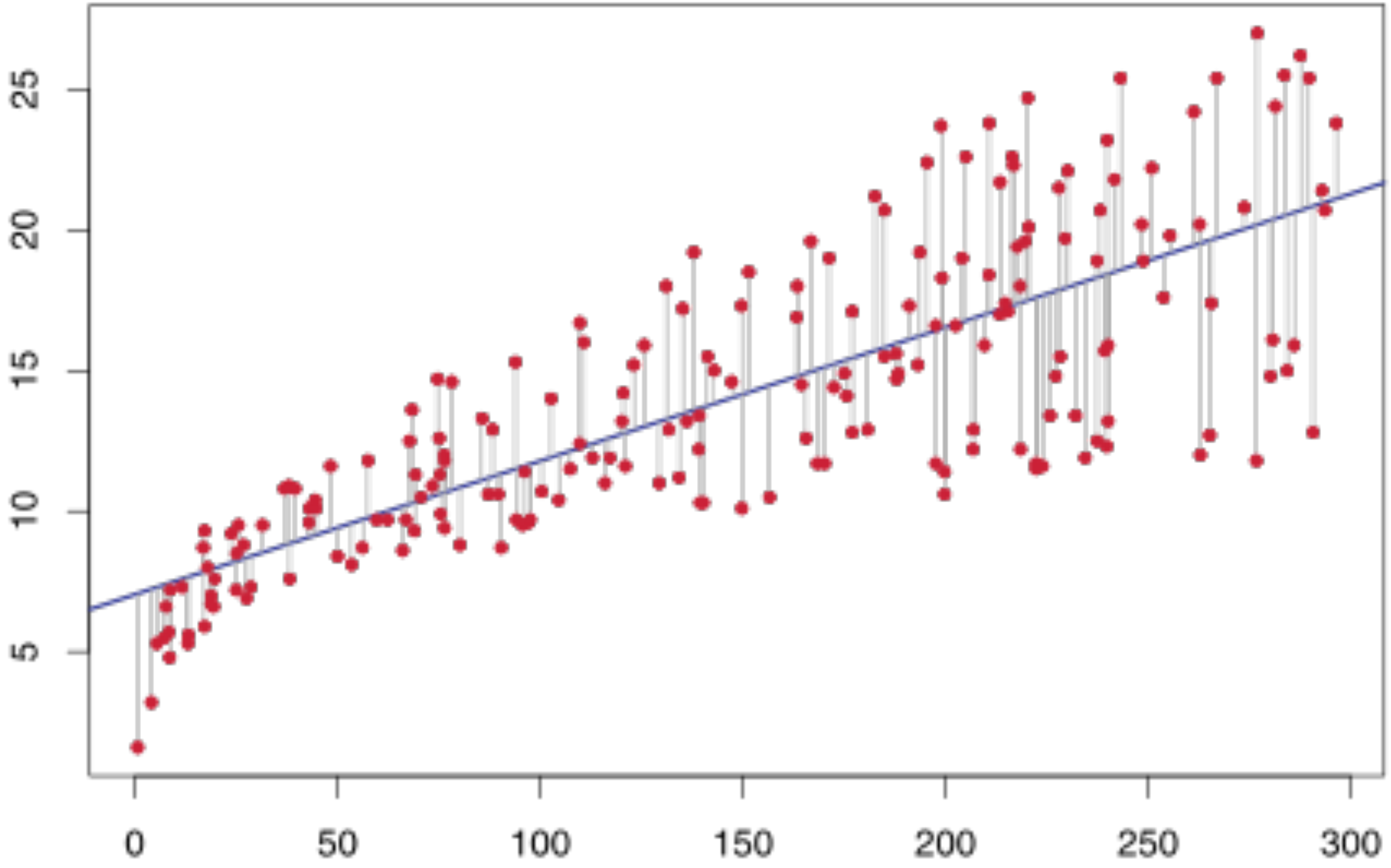
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i.$$

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i.$$

$$\sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2,$$



LINEAR REGRESSION



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

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

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

REGULARIZATION

Recall our earlier discussion of **overfitting.**

Recall our earlier discussion of **overfitting**.

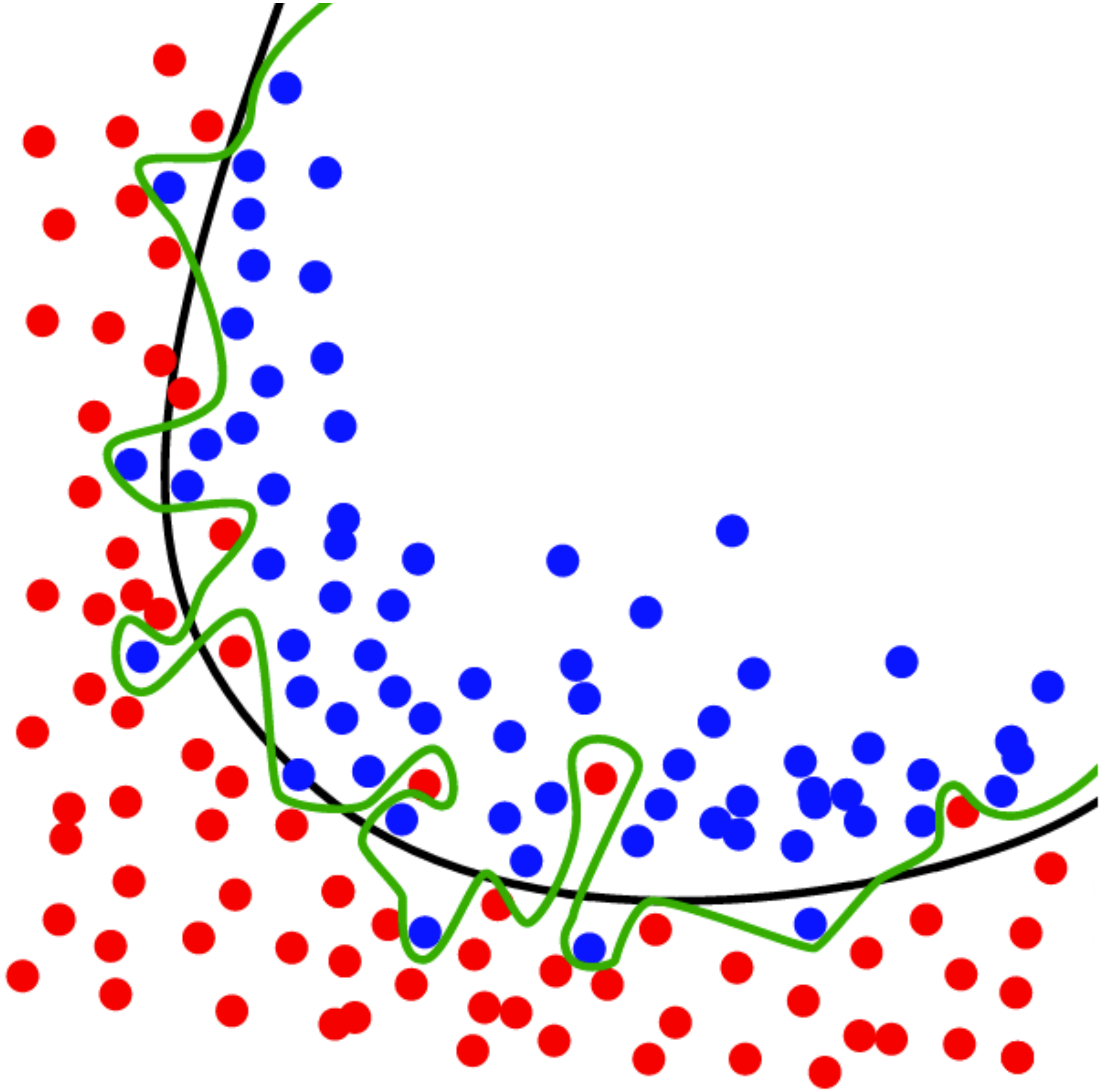
When we talked about this in the context of classification, we said that it was a result of matching the training set too closely.

Recall our earlier discussion of **overfitting***.*

When we talked about this in the context of classification, we said that it was a result of matching the training set too closely.

In other words, an overfit model matches the **noise** *in the dataset instead of the* **signal***.*

OVERFITTING EXAMPLE (CLASSIFICATION)



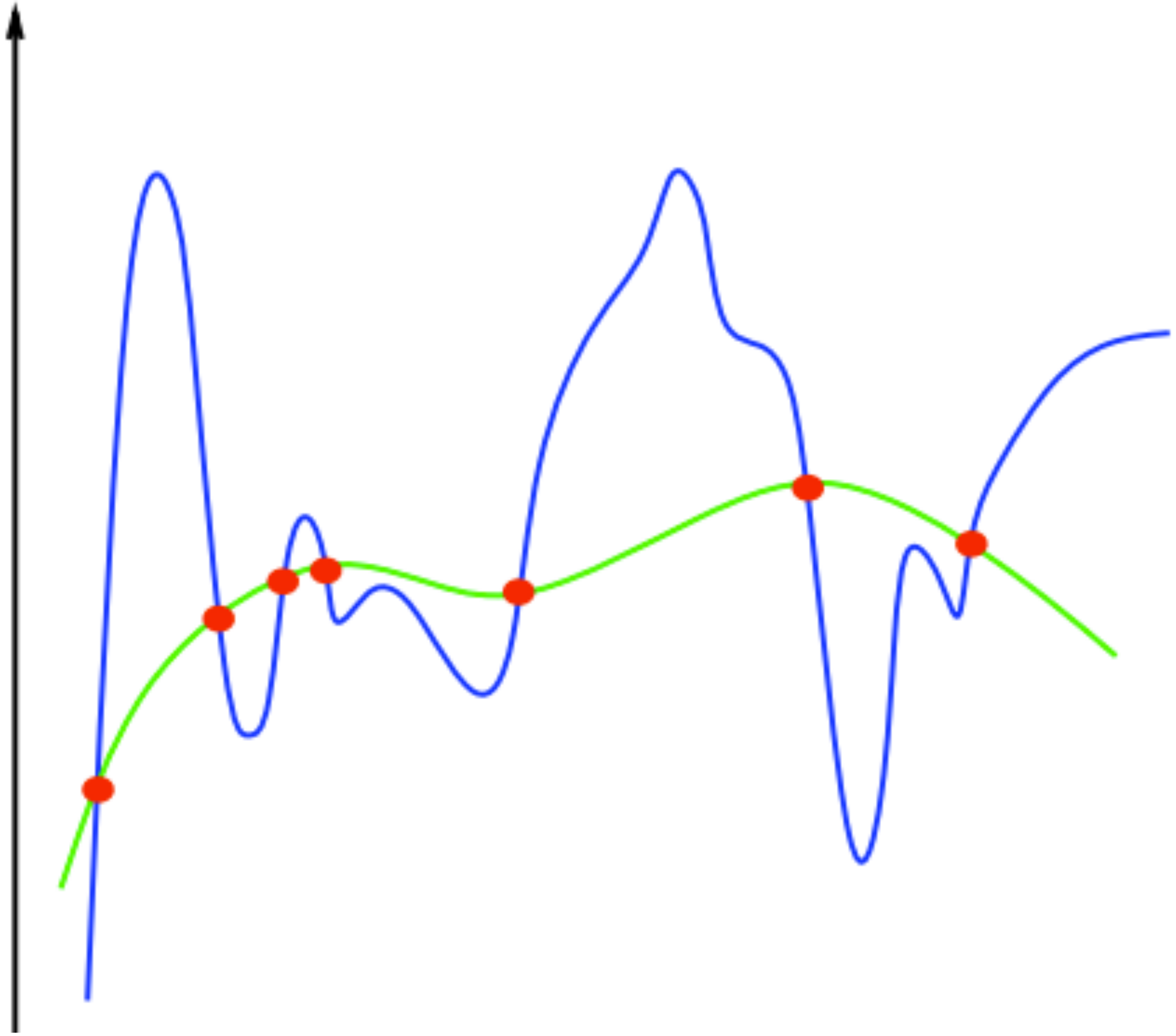
source: http://upload.wikimedia.org/wikipedia/commons/1/19/Overfitting.svg

The same thing can happen in regression.

It's possible to design a regression model that matches the noise in the data instead of the signal.

This happens when our model becomes too complex for the data to support.

OVERFITTING EXAMPLE (REGRESSION)



source: http://www.mit.edu/~9.520/spring12/slides/class02/class02.pdf

HOW DO WE SCREW UP AND OVERFIT?

- Testing and training on the same data

- Creating a model that is overly complex and only applies to the training data

- In other words, an overly complex model that can't generalize to new data

BUT WAIT, I THOUGHT LINEAR REGRESSION WAS A SIMPLE MODEL?

- Thought exercise: I add hundreds of irrelevant features to a model
- Linear regression will produce a coefficient for EACH of the hundreds
- This will produce a complex model that fits to the noise rather than the signal
- This is especially a problem when we have more features than observations

DID YOU MAKE SURE NONE OF YOUR FEATURES IS CORRELATED?

- If your features are correlated, your model will react to random errors
- In other words, random variations in the data will skew your model
- This means we're following the noise
- Overfitting!

HOW LARGE ARE YOUR COEFFICIENTS?

- If your coefficient is large, it has a ton of power to change the prediction
- This means that coefficient can cause huge fluctuations in your model
- In other words, if that coefficient has any noise, we could miss our target!

WE REGULARIZE THE SIZE OF THE COEFFICIENTS

- We tie our measure of model error to the size of the coefficients
- If our coefficients are large, we penalize the performance of our model
- This means coefficients need to balance actually reducing error with overfit
- This keeps an irrelevant feature from dominating our model
- Each feature needs to give its "fair share" to making our model effective!

RIDGE AND LASSO REGRESSION

- Ridge regression shrinks coefficients close to zero but never all the way
- This means that a coefficient of very close to zero is not very relevant in reducing error
- Lasso regressions shrinks irrelevant features all the way to zero

These regularization problems can also be expressed as:

L1 regularization (Lasso): $min(\|y - x\beta\|^2 + \lambda\|x\|)$

L2 regularization (Ridge): $min(\|y - x\beta\|^2 + \lambda\|x\|^2)$

This (Lagrangian) formulation reflects the fact that there is a cost associated with regularization.

THE TERMS NO ONE EXPLAINS

WHAT DOES SPARSE MEAN?

$$\begin{pmatrix} 11 & 22 & 0 & 0 & 0 & 0 & 0 \\ 0 & 33 & 44 & 0 & 0 & 0 & 0 \\ 0 & 0 & 55 & 66 & 77 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 88 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 99 \end{pmatrix}$$

sparse means only a subset of features are important -

we have zero'd out irrelevant features!

THERE IS NO FREE LUNCH

- You need to test performance!
- Research suggests but does not prove that ridge outperforms lasso

LASSO INHERENTLY PROVIDES FEATURE SELECTION!

- By pushing coefficients all the way to zero, lasso tells us what we can drop!
- This is a great way for us to make our model easier to explain to others

I HEARD OF INTERACTION EFFECTS…WHAT ARE THEY?

COMBO OF 2 OR MORE VARIABLES HAS A DIFFERENT EFFECT THAN ALONE

- Easiest to understand through examples:

- Adding sugar to coffee & stirring coffee -> sweetness
- Smoking & inhaling asbestos -> lung cancer

Consider your audience…

WHILE EXTREMELY USEFUL FOR ACCURACY, MAKES YOUR MODEL MORE COMPLEX

THERE IS NO FREE LUNCH

- You need to test performance!
- Research suggests but does not prove that regularization tends to outperform plain old linear regression, especially when high features / low observations
- Research suggests but does not prove that Ridge tends to outperform Lasso in terms of predictive performance

ASSUMPTIONS WE CARE ABOUT FOR PLAIN OLD LINEAR REGRESSION

- Linear relation between features and y (this is very rare in practice)
- Errors are independent (don't depend on changing x)
- Outliers and influential points are dangerous!

EVALUATING THE STRENGTHS AND WEAKNESSES OF LR

ADVANTAGES

- Easy to explain (relative)
- Parametric (we don’t need all the data to predict)
- Accurate if assumptions are met (otherwise performance is not great)

DISADVANTAGES

- Extremely sensitive to irrelevant features
- Fairly hard to meet assumptions in the real world
- Unable to automatically learn interactions among features

LET'S CODE!