GENERAL ASSEMBLY

INTRO TO DATA SCIENCE

LECTURE 5: THE LINEAR REGRESSION

LAST TIME:

- INTRO TO DATABASES
- WORKING WITH APIS AND JSON
- MYSQL QUERIES

QUESTIONS?

I. INTRODUCTION TO REGRESSION DATA PROBLEMS

II. HOW REGRESSIONS WORK

III. DETERMINING COST

EXERCISES:

IV. IMPLEMENTING THE LINEAR MODEL

I. LINEAR REGRESSION

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

	continuous	categorical
supervised	regression	classification
unsupervised	dimension reduction	clustering

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

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

A: A functional relationship between input & response variables.

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.

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)

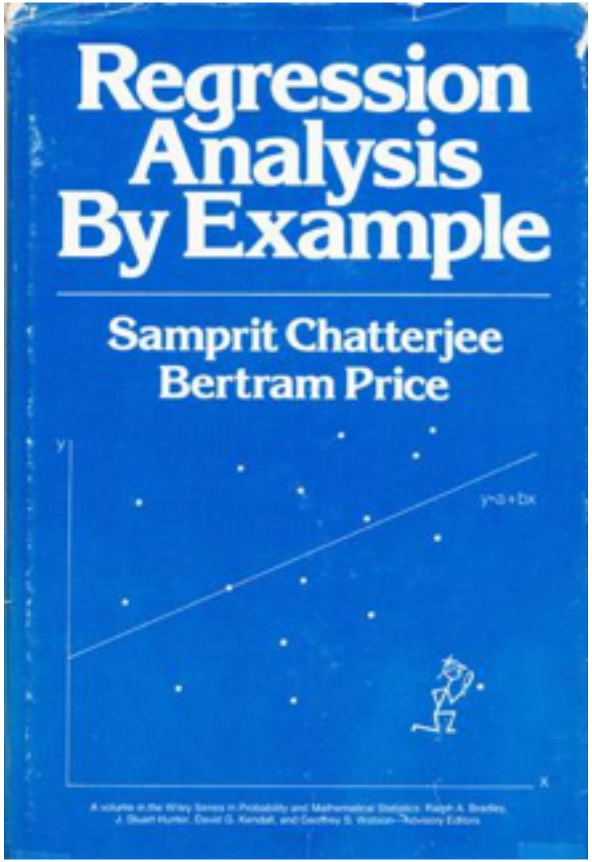
β = regression coefficient (the model "parameter")

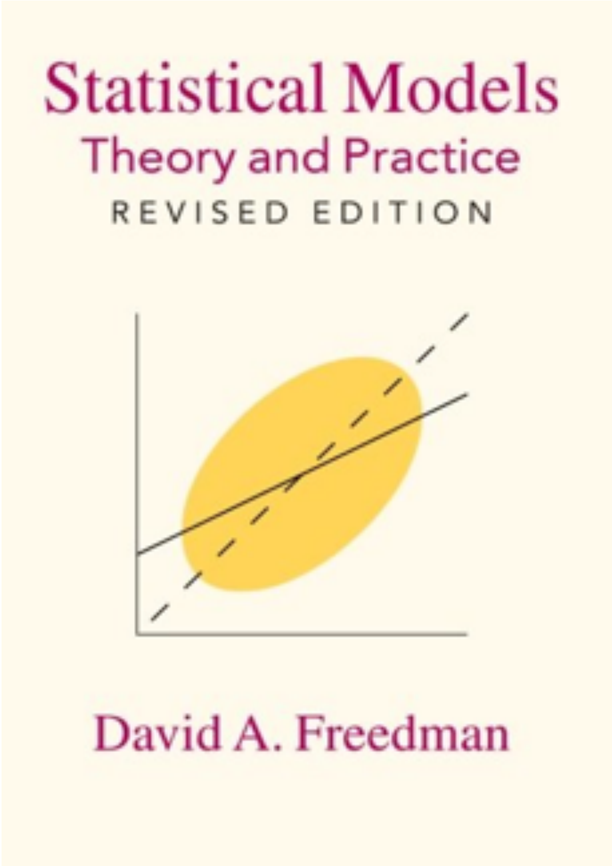
ε = residual (the prediction error)

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:

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


Regression Analysis By Example
Samprit Chatterjee
Bertram Price


Statistical Models
Theory and Practice
REVISED EDITION
David A. Freedman

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

In order for us to gain a deeper understanding of the "magic" behind a regression (and to see why we want a machine to do this!), let's review the math behind this algorithm!

II: THE MATH WAY

Linear regression is, for the most part, just matrix algebra (the stuff we did already!)

Let's go over the math by hand so we can understand how we determine the **regression coefficient**.

A linear regression in its simplest form:

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

A linear regression in its simplest form:

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

but we can assume that our α is either 0 or 1, and ε is zero!

$$y = \beta x$$

So in a more simple form:

$$y = \beta x$$

but we want to solve for β, which means our new equation looks more like this:

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

So if we had data:

3.385	44.5
0.48	15.5
1.35	8.1
465	423
36.33	119.5

So if we had data:

3.385	44.5
0.48	15.5
1.35	8.1
465	423
36.33	119.5

Response

Input

$$\left(\begin{matrix} 1 & 1 & 1 & 1 & 1 \\ 3.385 & 0.48 & 1.35 & 465 & 36.33 \end{matrix} \begin{matrix} 3.385 & 1 \\ 0.48 & 1 \\ 1.35 & 1 \\ 465 & 1 \\ 36.33 & 1 \end{matrix} \right)^{-1}$$

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

$$\left(\begin{array}{ccccc} 1 & 1 & 1 & 1 & 1 \\ 3.385 & 0.48 & 1.35 & 465 & 36.33 \end{array} \begin{array}{c} 44.5 \\ 15.5 \\ 8.1 \\ 423 \\ 119.5 \end{array} \right)$$

... $X^T y$

$$\begin{matrix} 0.2617 & -0.0006 \\ -0.0006 & 0.000006 \end{matrix} \qquad \begin{matrix} 610.6 \\ 201205.4425 \end{matrix}$$

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

$$\begin{matrix} 37.2 \\ 0.838 \end{matrix} = \begin{matrix} 0.2617 & -0.0006 \\ -0.0006 & 0.000006 \end{matrix} \quad \begin{matrix} 610.6 \\ 201205.4425 \end{matrix}$$

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

Intercept

$$\begin{bmatrix} 37.2 \\ 0.838 \end{bmatrix} = \begin{bmatrix} 0.2617 & -0.0006 \\ -0.0006 & 0.000006 \end{bmatrix} \begin{bmatrix} 610.6 \\ 201205.4425 \end{bmatrix}$$

β

$$\beta = (X^TX)^{-1} X^Ty$$

Q: How did we do compared to a computer?

Q: How did we do compared to a computer?

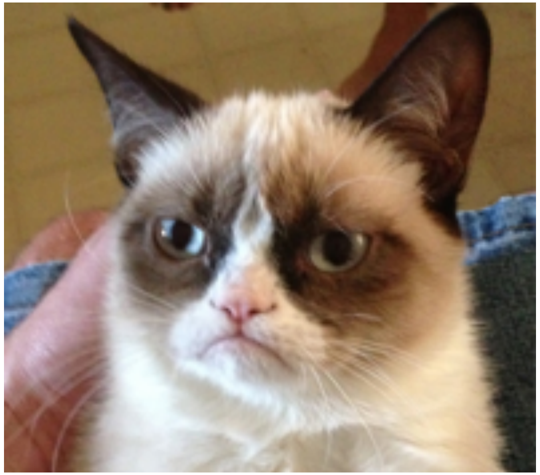
```
Call:
lm(formula = brain ~ body, data = head(mammals, 5))

Coefficients:
(Intercept)         body
    37.2009       0.8382
```

A: Not bad!

Q: Cool! That means we can do all of our regressions by hand now, right?

Q: Cool! That means we can do all of our regressions by hand now, right?



REVIEW: MATRIX ALGEBRA

Review this concept with data that we know has a coefficient of 1 and an intercept of 0:

Input		Response
1	1	
2	2	
3	3	
4	4	

III: COST OF LINEAR REGRESSIONS

Q: How do measure error in a linear regression model?

Q: How do measure error in a linear regression model?

A: In theory, **minimize the sum of the squared residuals (RSS, or SSE).**

Q: How do measure error in a linear regression model?

A: In theory, **minimize the sum of the squared residuals (RSS, or SSE).**

In practice, any respectable software can do this for you.

Q: How do measure error in a linear regression model?

A: In theory, **minimize the sum of the squared residuals (RSS, or SSE).**

In practice, any respectable software can do this for you.

In python, we can find this with some quick code.

Q: How do measure error in a linear regression model?

A: In theory, **minimize the sum of the squared residuals (RSS, or SSE).**

In python, we can find this with some quick code:

mean((prediction – actual)^2)

Q: How do measure goodness of fit?

A: In theory, we want to **maximize R^2 (as close to one as possible).**

Q: How do measure goodness of fit?

A: In theory, we want to **maximize R^2 (as close to one as possible).**

Scikit Learn already calculates this for us, as do any other stats packages and programs.

Q: How do measure goodness of fit?
A: In theory, we want to **maximize R^2 (as close to one as possible).**

Scikit Learn already calculates this for us, as do any other stats packages and programs.

If you want to get serious into regression, learn more about the **coefficient of determination**.

REVIEW: COST

1. What values are we looking for when we consider **SSE**? What is the best value we could potentially have?

2. What is the best value we could have for R^2?

3. What's the primary difference between these two values?

EX: LINEAR REGRESSIONS

NEXT TIME:
POLYNOMIAL AND LOGISTIC REGRESSIONS