Lecture 4 –EDA Fundamentals Continued

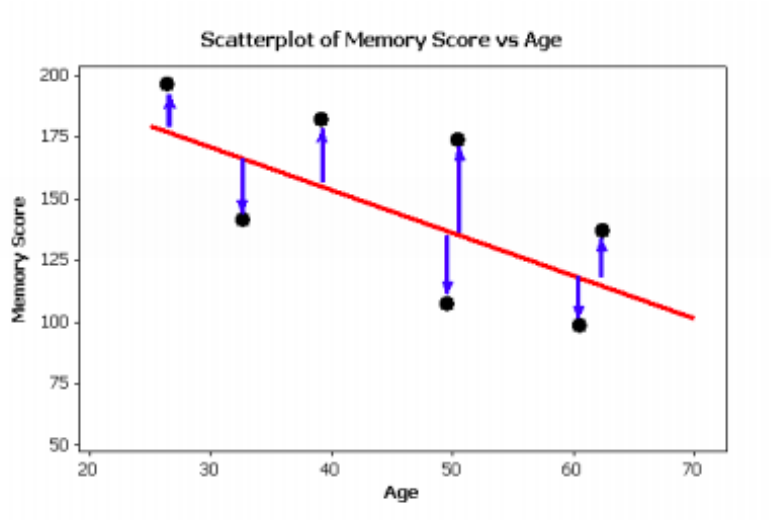
Determining the line that best fits the data

Goal - predicting the response based on the explanatory variable(s) & quantifying the linear effect by interpreting the slope (and intercept) of the line **in context**

Motivating Example – Does age impact memory score?

How is this line quantifiably constructed?

1. Theoretically



2. Formulaically

A word of caution: the regression line is heavily influenced by outliers. This is why it's so important to identify them first. Think of the outlier pulling the line toward it.

3. Example – Does cost of a car impact the maximum number of miles it can go before breaking down?

   Dependent:
   Independent:

   Descriptive Statistics:

	n	$\bar{x}$	s	min	max	Q1	Q3	M
Cost	235	\$34,500	\$6,500	\$11,200	\$112,000	\$23,600	\$67,800	\$56,000
Miles	235	98,000	4,500	74,000	130,000	86,000	112,000	95,000

and after running the initial simple linear regression in Minitab, we get a correlation coefficient of $r = 0.78$

What is the regression equation?

How to use the regression line

1. Prediction –

Example – Car cost impacting total miles drive example continued

How many miles, on average, will a car that costs $92,000 go?

What about a car that costs $17,000?

2. Quantifying the linear effect - interpreting the slope

Let's revisit at the car cost example. The regression equation was

Now, how many miles will a car that costs $81,000 go before breaking down?

How many miles will a car that costs $81,001 go before breaking down?

What is the difference? Notice anything?

The slope of the regression line represents the unit change, on average.

Slope interpretation template:

For every one-unit increase in the explanatory variable, there is, on average, a (slope) increase/decrease in the response variable.

So, back to the car cost example, we would interpret the slope as follows:

What about the intercept?

What does it imply in the context of the car cost example?

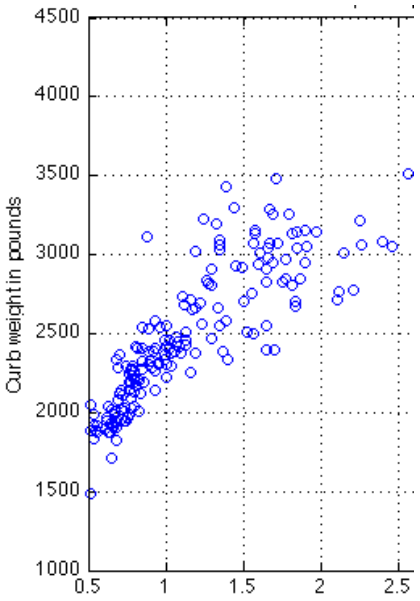
This will not always be the case. Depending on the context, the intercept can be very useful, or completely irrelevant. This is where you come in as a statistician.

Beware of the temptation to extrapolate

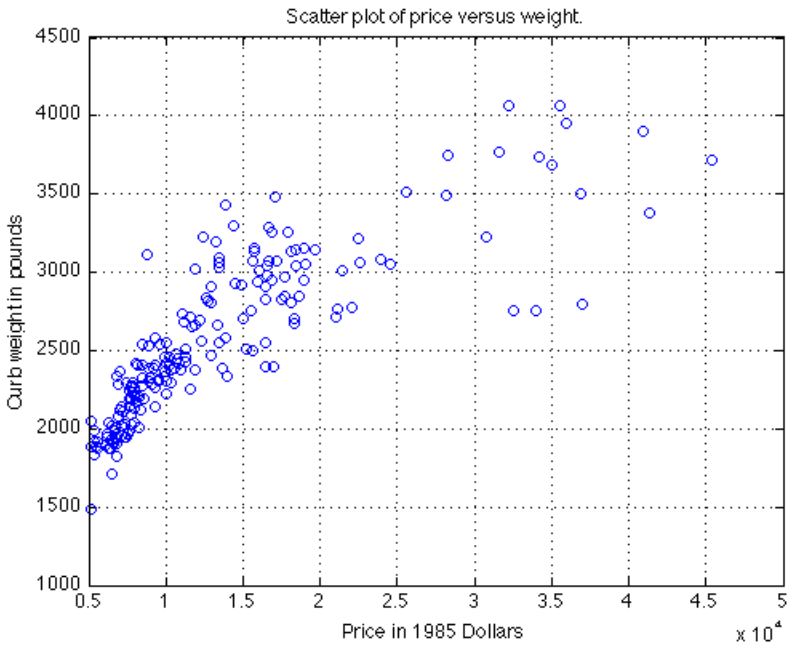
There is no way to know if the relationship holds outside of the range of the explanatory variable that we have data for.

Let's consider an example…

This is a scatterplot of price by weight. Consider this range in price from $0.5 to $2.5 in 1985 US dollars. What would the regression line look like? What would you predict the weight to be for $4.0 1985 US dollars?



Now consider the entire scatterplot that we have information for. What would the regression line look like? Now what would you predict the weight to be for $4.0 1985 US dollars?



This is why you don't extrapolate!

Quantifying Prediction Error …the line is not perfect!

Let's look at some Minitab output when running a simple linear regression:

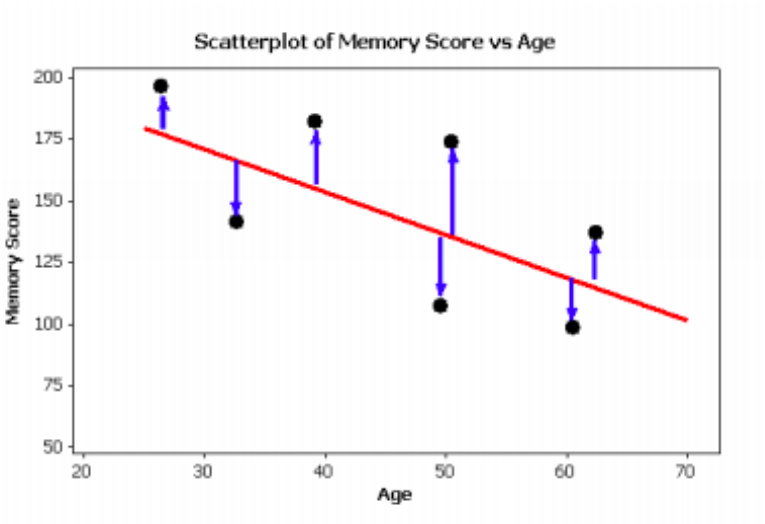
Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
19.1942	53.36%	49.78%	37.02%

Regression Equation

Mem = 218.3 - 1.698 Age

a. Standard error of the regression (s) –



Let's consider the example examining the impact of age on memory score:

	Age	Memory Score	Fitted Values (y-hat)	Residuals	Residuals^2
1	33	145	165.671	-20.6710	427.29
2	54	112	129.237	-17.2369	297.11
3	50	172	136.177	35.8232	1283.30
4	28	158	174.346	-16.3458	267.19
5	42	158	150.056	7.9436	63.10
6	34	166	163.936	2.0639	4.26
7	47	159	141.382	17.6184	310.41
8	46	121	143.117	-22.1166	489.14
9	65	109	110.152	-1.1524	1.33
10	46	136	143.117	-7.1166	50.65
11	60	130	118.827	11.1728	124.83
12	64	78	111.887	-33.8874	1148.35
13	41	163	151.791	11.2086	125.63
14	36	143	160.466	-17.4662	305.07
15	33	171	165.671	5.3290	28.40

b. Coefficient of Variation $R^2 = r^2$ =

i. $0 \leq R^2 \leq 1$

ii. Use for model comparison ONLY → choose the model with the highest R^2

iii. Think about what it means to take R^2 at one of its boundary points

iv. Can go from r to R^2, but careful going from R^2 to r → have to look at the scatterplot to get the direction (sign of the correlation coefficient r)

v. To summarize, we want the model with the largest R^2, or equivalently the model with the largest |r|, or the model with the smallest s (all imply "tighter" or stronger relationship)

A warning: Don't dismiss small values of R^2. It's a comparative measure, so if the best model we have still has a low R^2, then we have no choice.

To summarize:

2 Quantitative Variables (Q → Q):

Graphical:

Numerical Summary:

ASSOCIATION DOES NOT IMPLY CAUSATION. REGRESSION ONLY ALLOWS US TO INFER ASSOCIATION, NOT CAUSATION.

Q → C Beyond the scope of this course