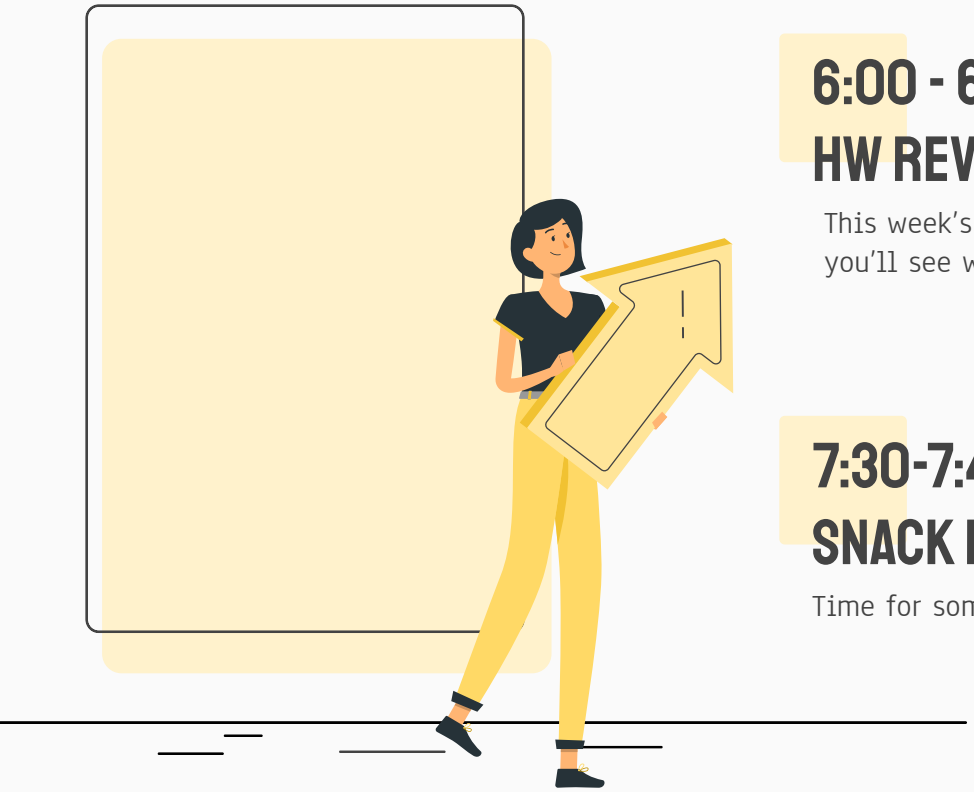
APPLIED DATA SCIENCE II

Week 2: LINEAR MODELS!

Kyle Scot Shank
WI-22



6:00 - 6:30
HW REVIEW

This week's a little different, you'll see why!

6:30-7:30
TOPIC OVERVIEW

This week, we're talking about all things *linear*

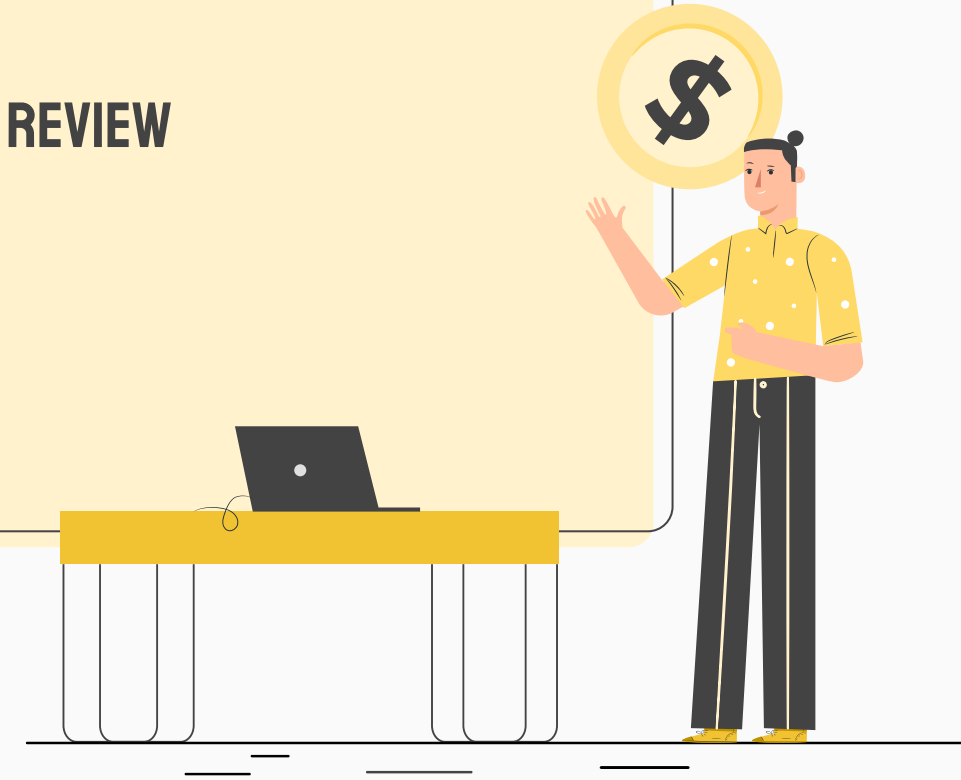
7:30-7:45
SNACK BREAK!

Time for some munchies

7:45 - 9:00
HANDS-ON CODE LAB

Work through stuff together

HW REVIEW



HW REVIEW

WHAT IS A STATISTICAL PREDICTION?

A statistical method that attempts to accurately project the chances that something will (or will not) happen

HW REVIEW

WHAT ARE THE BASIC COMPONENTS OF ANY PREDICTIVE MODEL?

Reducible error and irreducible error.

HW REVIEW

WHAT IS A FLEXIBLE VS. AN INFLEXIBLE PREDICTIVE MODEL?

*A **flexible** model is one that can take many different functional forms, but might have less interpretability.*

*An **inflexible** model is one that has a single/specific functional form, but is often highly interpretible.*

HW REVIEW

WHAT'S THE DIFFERENCE BETWEEN INFERENCE AND PREDICTION?

Prediction *is focused on estimating future outcomes as accurately as possible.*

Inference *is focused on understanding the present mechanisms of action as accurately as possible.*

HW REVIEW

WHAT'S THE DIFFERENCE BETWEEN REGRESSION AND CLASSIFICATION?

Regression models *are focused on predicting a quantitative response variable (i.e., a number)*

Classification models *are focused on predicting is focused on understanding the present mechanisms of action as accurately as possible.*

HW REVIEW

DESCRIBE THE VARIANCE-BIAS TRADEOFF IN SIMPLE WORDS

Bias *is the simplifying assumptions made by the model to make the target function easier to approximate.*

Variance *is the amount that the estimate of the target function will change given different training data.*

In general, as you reduce bias - you increase variance, and vice versa*.*

HW REVIEW

WHAT IS THE MEAN SQUARED ERROR USED FOR?

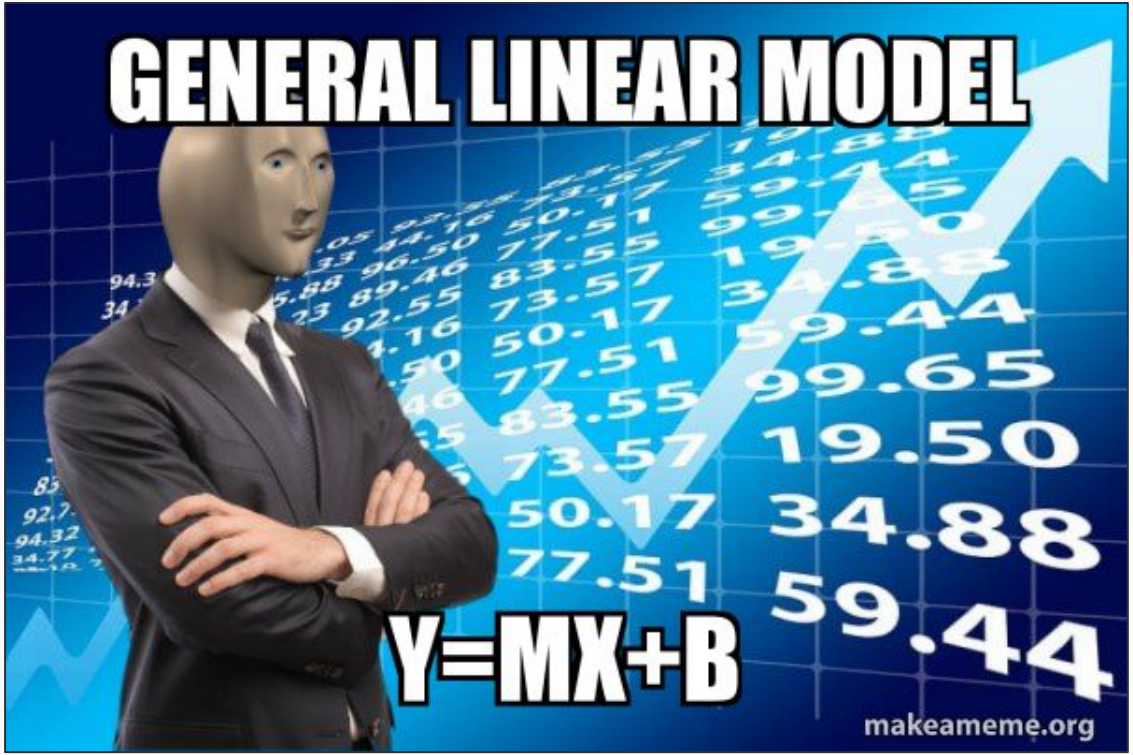
*To measure the quality of fit (**which is equivalent to measuring the accuracy of predictions for a linear model!**)*

HW REVIEW

RMARKDOWN TROUBLESHOOTING!

TOPIC OVERVIEW

LINEAR MODELS!



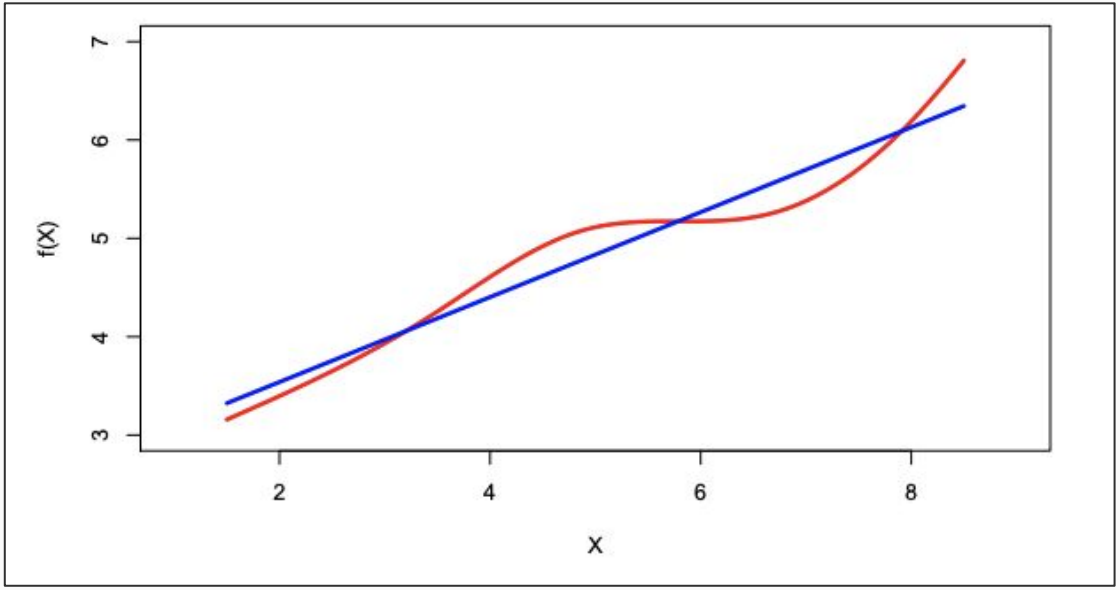
WE'RE GOING TO TALK ABOUT THE GOOD OL' BREAD AND BUTTER OF EVERY DATA SCIENTIST: LINEAR MODELS

LINEAR REGRESSION

Linear regression is a simplest approach to supervised learning. It assumes that the dependence of Y on X_1, X_2, . . . X_p is **linear**.

That said, true relationships are almost **never** *linear!*

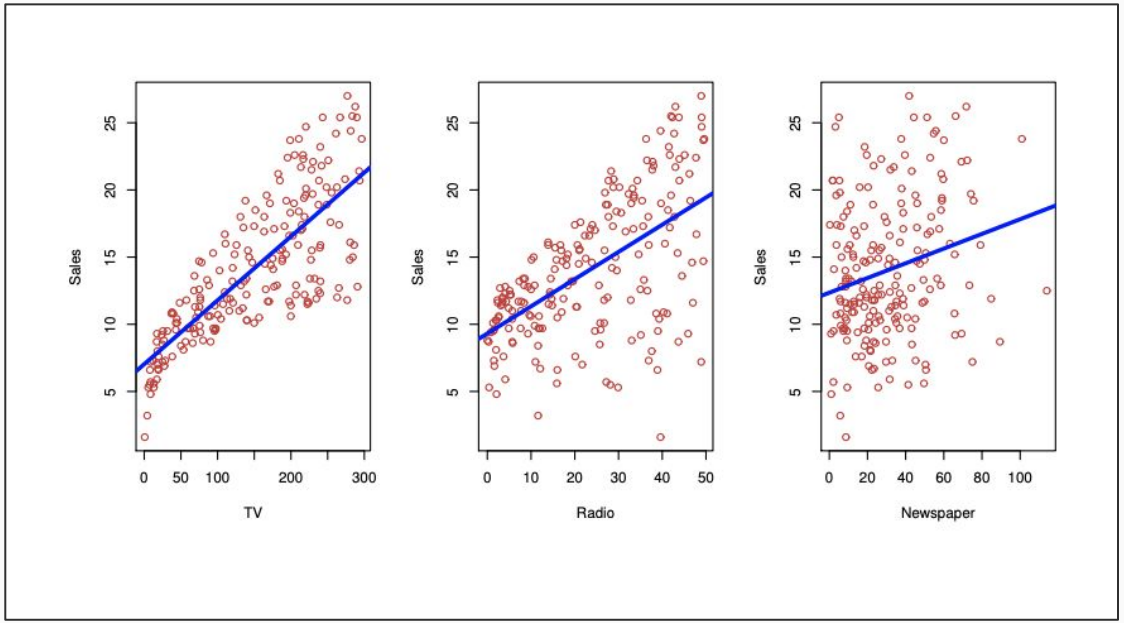
...but approximating them in a linear way makes things really, really convenient



LINEAR REGRESSION

Consider the advertising data shown here. Questions we might ask:

- *Is there a relationship between advertising budget and sales?*
- *How strong is the relationship between advertising budget and sales?*
- *Which media contribute to sales?*
- *How accurately can we predict future sales?*
- *Is the relationship linear?*



LINEAR REGRESSION

First things first! *We assume a model of the following form:*

$$Y = \beta 0 + \beta 1X + \varepsilon,$$

where $\beta 0$ *and* $\beta 1$ *are two unknown constants that represent the intercept and slope, also known as coefficients or parameters, and* ε *is the error term.*

Given some estimates $\hat{\beta} 0$ *and* $\hat{\beta} 1$ *for the model coefficients, we predict future sales using*

$$\hat{y} = \hat{\beta} 0 + \hat{\beta} 1x$$

where $\hat{y}$ *indicates a prediction of Y on the basis of X = x. The hat symbol denotes an estimated value.*

...now, let's actually estimate these parameters!

Let

$$\hat{y}_i = \hat{\beta}0 + \hat{\beta}\,1x_i$$

be the prediction for Y based on the ith value of X. Then

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

represents the ith residual (i.e., error between predicted and actual values)

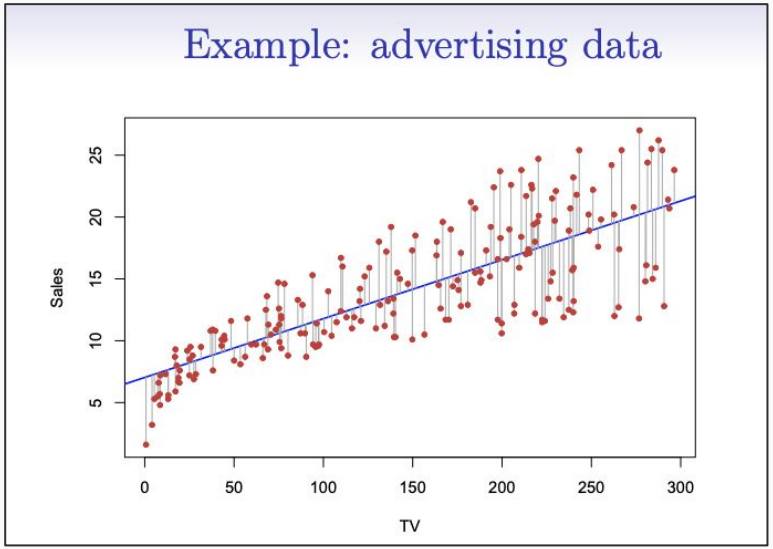
We then define the residual sum of squares (RSS) as

$$RSS = e2_1 + e2_2 + \cdots + e2_n$$

LINEAR REGRESSION

What the least squares approach does is minimizes the value of RSS - hence "least squares".

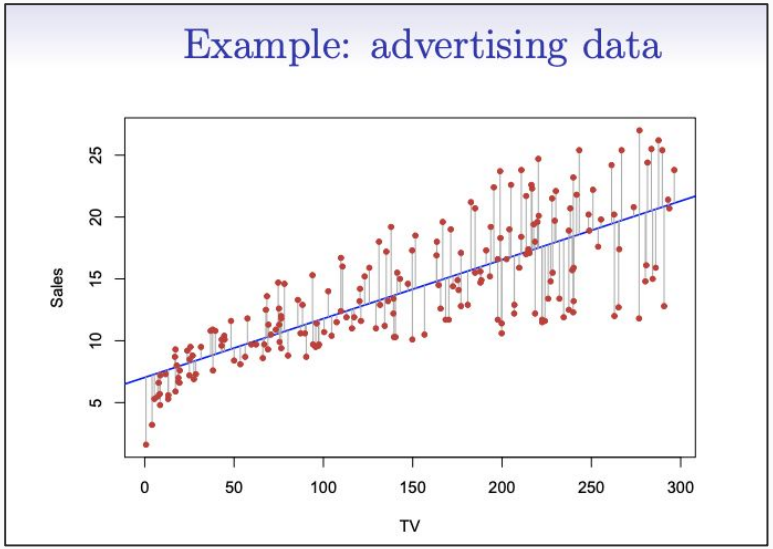
On the right you can see the least squares fit for the regression of the dependent variable sales onto the independent variable TV. In this case a linear fit captures the essence of the relationship, although it is somewhat deficient in the left of the plot.



After we've generated our estimated coefficient, we can then assess our confidence in it.

I won't walk through the calculation here - but we build a confidence interval out of the standard errors to say how confident we are in the true value being "close" to the estimated value of our coefficient. In general, we use 95% confidence intervals.

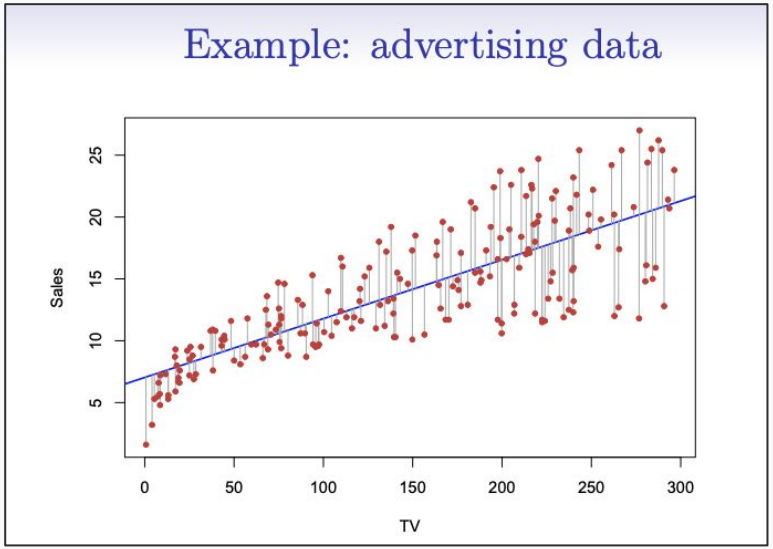
For the TV coefficient data, the 95% confidence interval for $\beta 1$ is ***[0.042, 0.053]***



After we've generated our estimated coefficient, we can then assess our confidence in it.

I won't walk through the calculation here - but we build a confidence interval out of the standard errors to say how confident we are in the true value being "close" to the estimated value of our coefficient. In general, we use 95% confidence intervals.

For the TV coefficient data, the 95% confidence interval for $\beta 1$ is ***[0.042, 0.053]***



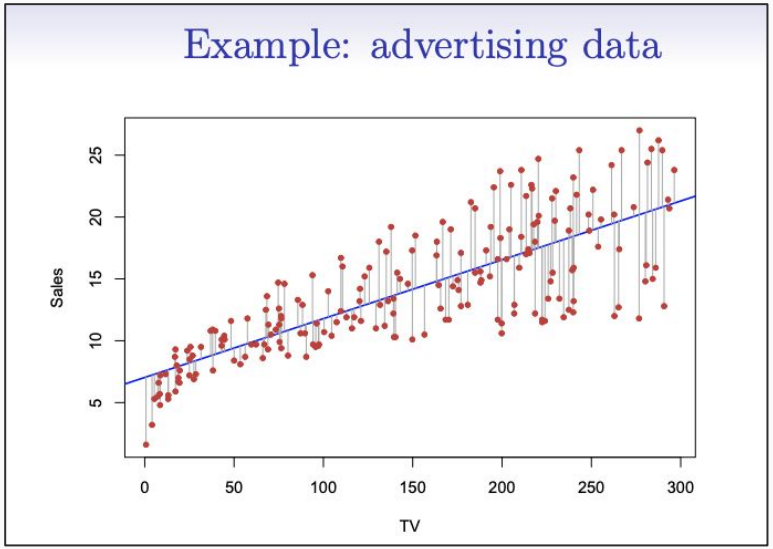
Why do you care about confidence intervals?
So we can run hypothesis tests!

Standard errors can be used to perform hypothesis tests on the coefficients. The most common hypothesis test involves testing the null hypothesis of

H0 : *There is no relationship between X and Y*

versus the alternative hypothesis

HA : *There is some relationship between X and Y .*



To test the null hypothesis, we compute a ***t-statistic****, given by*

$t = (\hat{\beta}_1 - 0)/SE(\hat{\beta}_1)$

This will have a t-distribution with n - 2 degrees of freedom, assuming $\beta_1 = 0$.

Using R, it is easy to compute the probability of observing any value equal to |t or larger. We call this probability the ***p-value****.*

	Coefficient	Std. Error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

Finally, with all of these values, we can then actually assess the performance of our model!

We'll skip the formulas, but understanding what these are is important:

Residual Standard Error *is a measure of how well the model fits the underlying dataset.*

R^2 *is a fractional measure of how much variance is explained by the model.*

F-statistic *measures how much better your model is vs. one with no variables*

Quantity	Value
Residual Standard Error	3.26
R^2	0.612
F-statistic	312.1

LINEAR REGRESSION

Question:

If you remember the reading from the Applied Predictive Modeling text - there's a line about why many people focus on ***root mean squared error*** *(RMSE) versus* ***residual standard error****. (also sometimes called the mean squared error).*

Do you remember why?

This is a hint!

Interpreting regression coefficients

- *Don't claim causality - claim correlation.*
- *Remember that the actual interpretation of any given coefficient is ""a unit change in X_i is associated with a βI change in Y , **while all the other variables stay fixed**".*
- *Making sure your predictors are uncorrelated is really important - we'll talk more about that.*

Interpreting predictions from a model

- *Don't claim causality - claim correlation.*
- *Remember that the actual interpretation of any given coefficient is ""a unit change in X_i is associated with a βI change in Y , **while all the other variables stay fixed**".*
- *Making sure your predictors are uncorrelated is really important - we'll talk more about that.*

SNACK BREAK!

COME BACK IN 15!

CODE LAB!

OPEN UP RSTUDIO