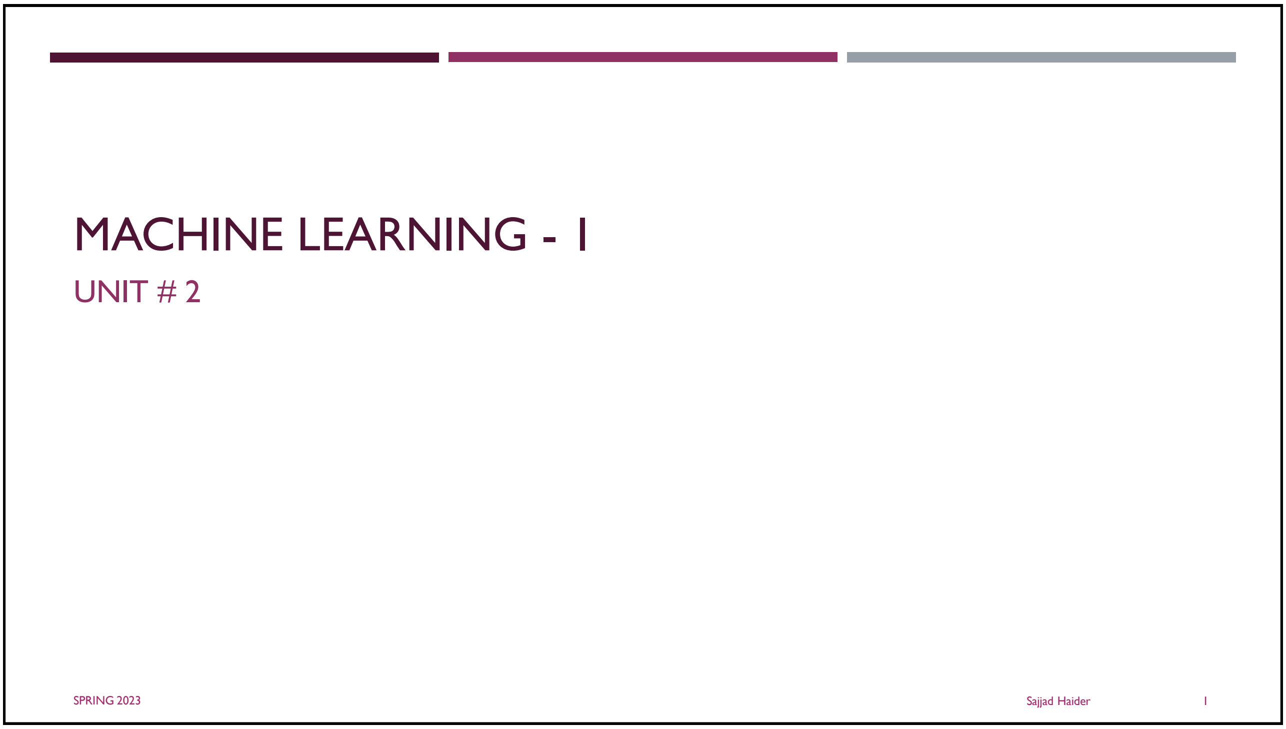
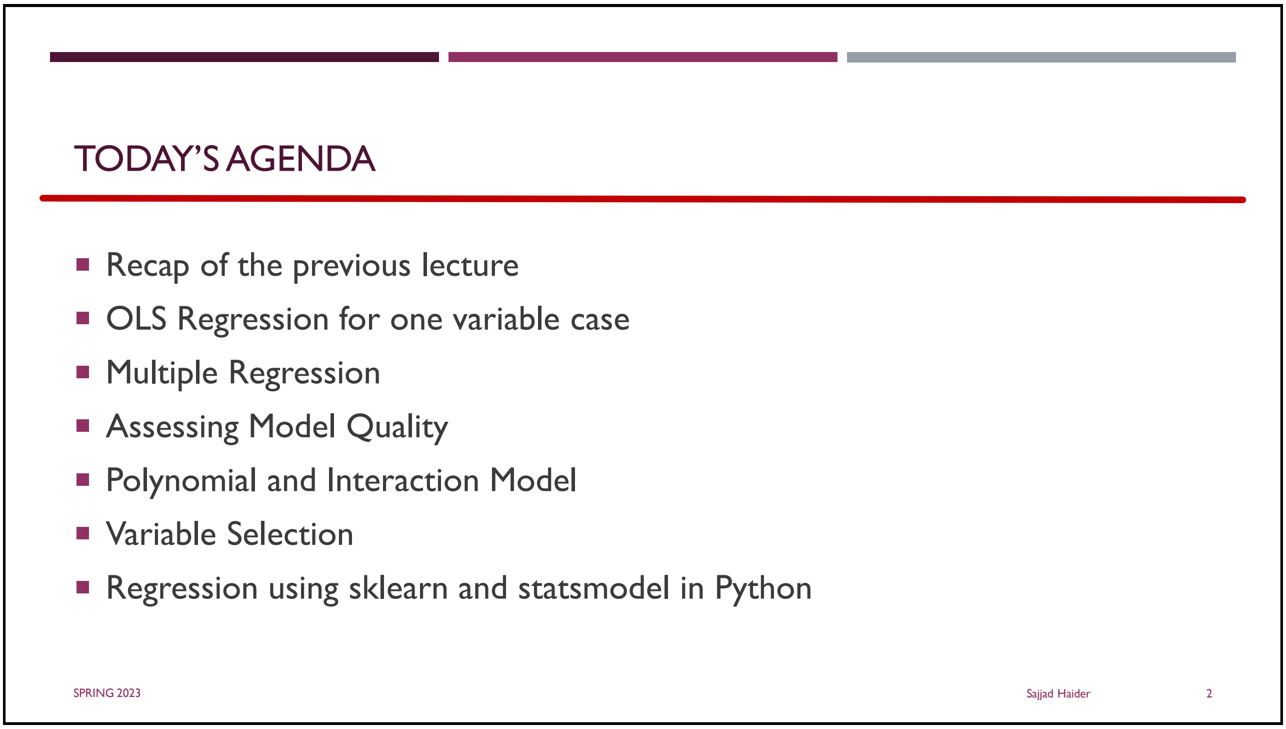


MACHINE LEARNING - I

UNIT # 2

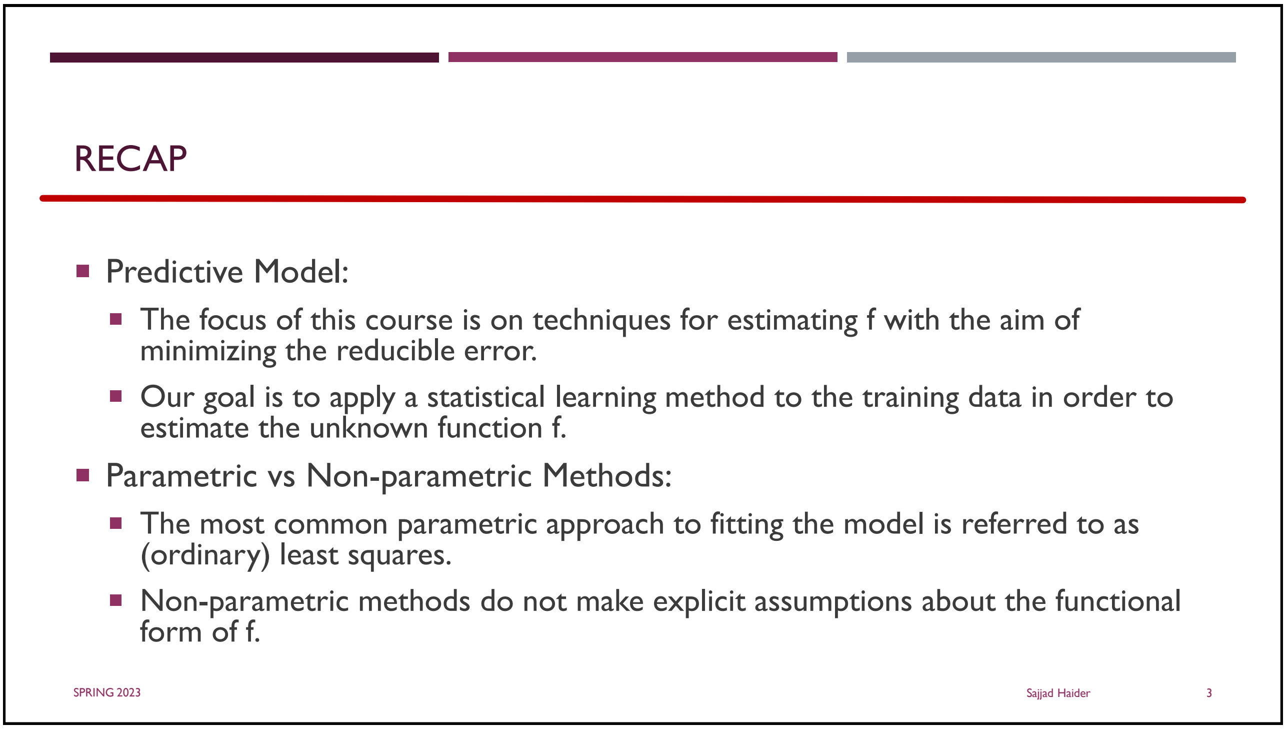
1



TODAY'S AGENDA

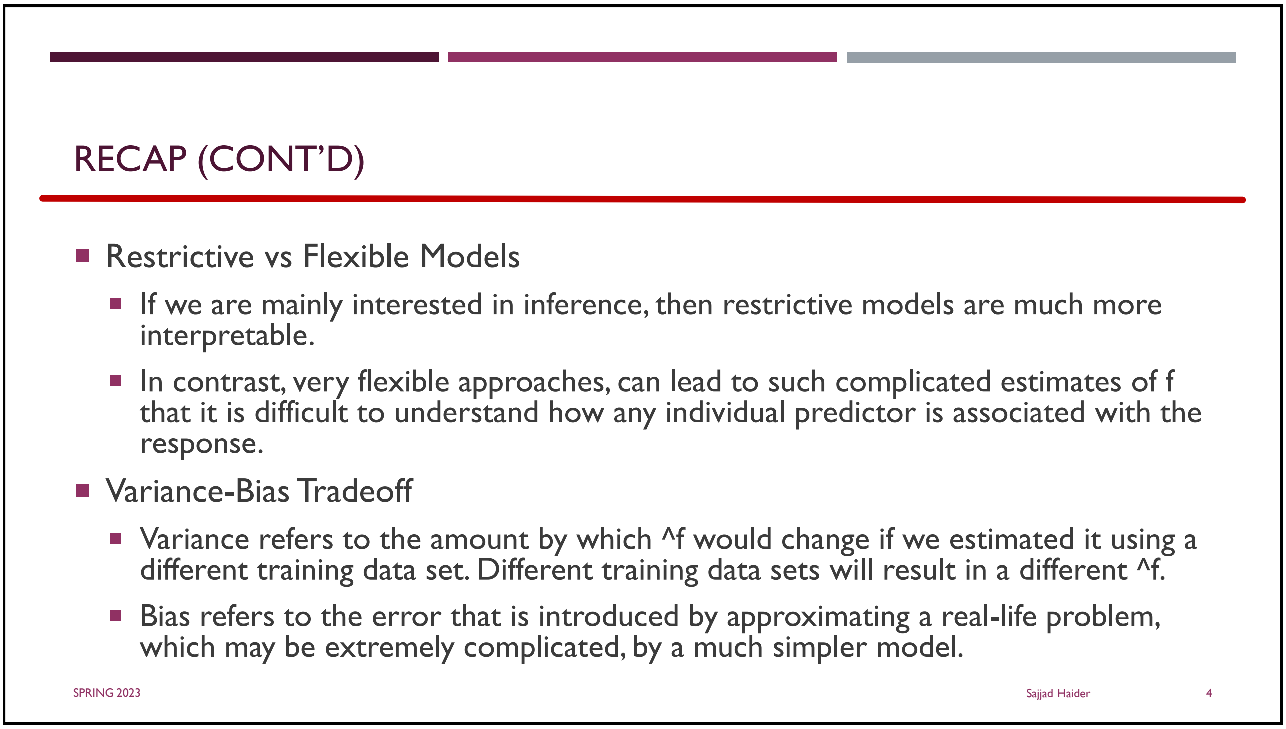
- Recap of the previous lecture
- OLS Regression for one variable case
- Multiple Regression
- Assessing Model Quality
- Polynomial and Interaction Model
- Variable Selection
- Regression using sklearn and statsmodel in Python

2



RECAP

- Predictive Model:
  - The focus of this course is on techniques for estimating f with the aim of minimizing the reducible error.
  - Our goal is to apply a statistical learning method to the training data in order to estimate the unknown function f.
- Parametric vs Non-parametric Methods:
  - The most common parametric approach to fitting the model is referred to as (ordinary) least squares.
  - Non-parametric methods do not make explicit assumptions about the functional form of f.



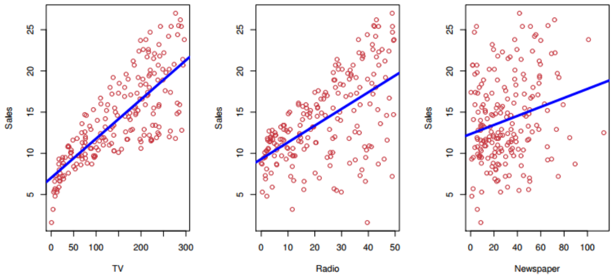
RECAP (CONT'D)

- Restrictive vs Flexible Models
  - If we are mainly interested in inference, then restrictive models are much more interpretable.
  - In contrast, very flexible approaches, can lead to such complicated estimates of f that it is difficult to understand how any individual predictor is associated with the response.
- Variance-Bias Tradeoff
  - Variance refers to the amount by which ^f would change if we estimated it using a different training data set. Different training data sets will result in a different ^f.
  - Bias refers to the error that is introduced by approximating a real-life problem, which may be extremely complicated, by a much simpler model.

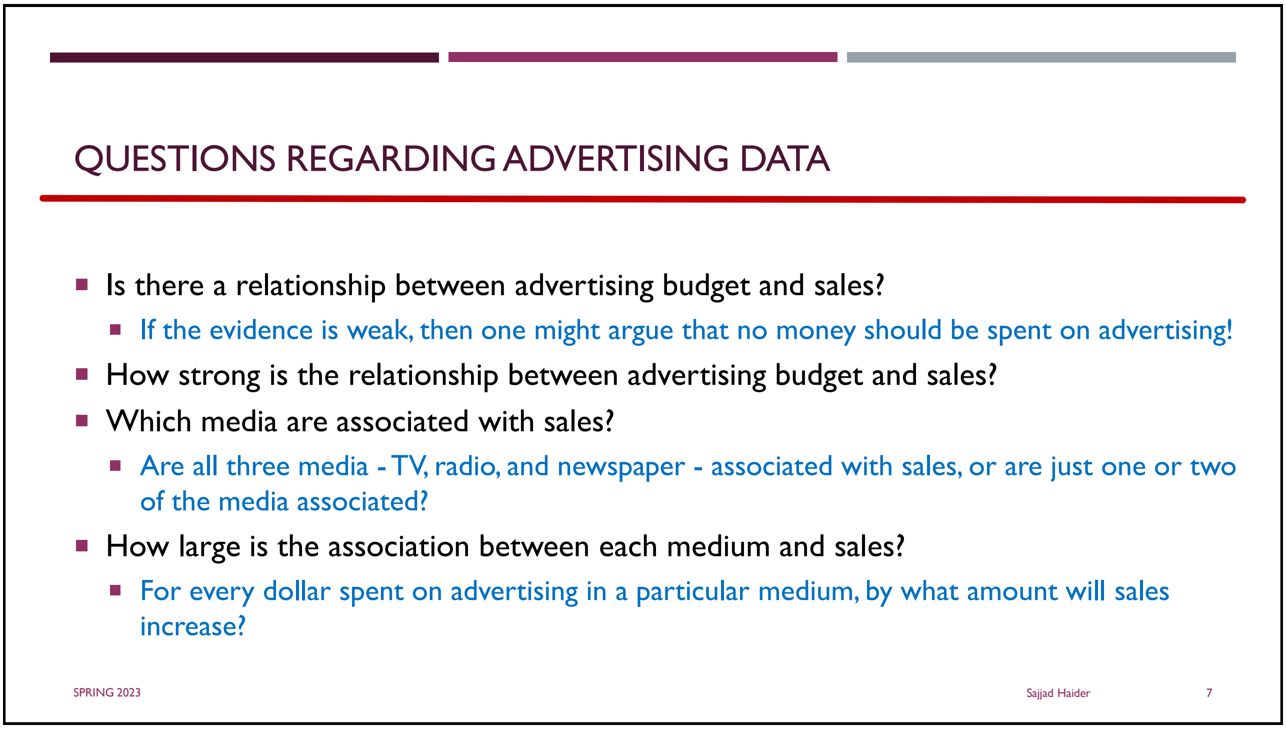
RECAP (CONT'D)

- **Overfitting** means a model fits the existing observations **too well** but fails to predict future new observations. This can occur when we're over extracting too much information from the training sets and making our model just work well with them, which is called **low bias** in machine learning. The model, as a result, will perform poorly on datasets that weren't seen before. We call this situation **high variance** in machine learning.
- The opposite scenario is **underfitting**. When a model is underfit, it doesn't perform well on the training sets and won't do so on the testing sets, which means it fails to capture the underlying trend of the data. We call any of these situations a high **bias** in machine learning; although its variance is low as the performance in training and test sets is pretty consistent, in a bad way.

SAMPLE DATA: ADVERTISING

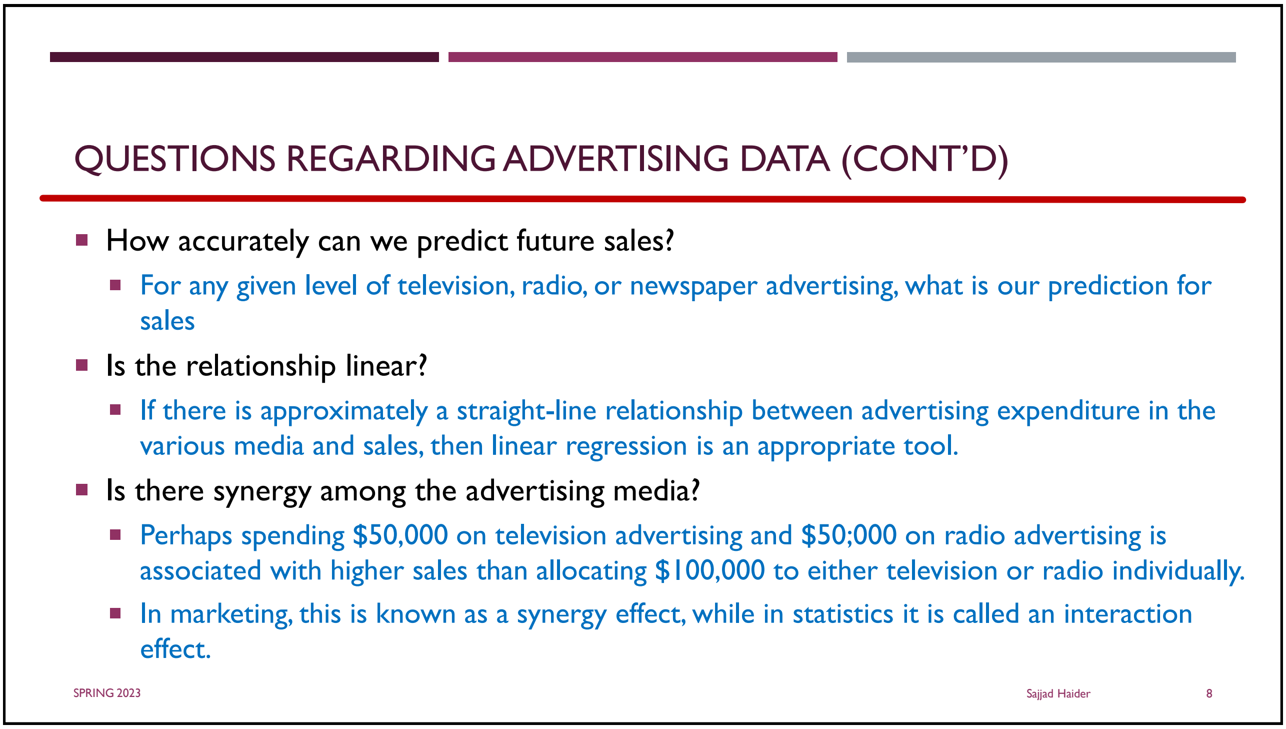


TV	Radio	Newspaper	Sales
230.1	37.8	69.2	22.1
44.5	39.3	45.1	10.4
17.2	45.9	69.3	9.3
151.5	41.3	58.5	18.5
180.8	10.8	58.4	12.9
8.7	48.9	75	7.2
57.5	32.8	23.5	11.8
120.2	19.6	11.6	13.2
8.6	2.1	1	4.8
199.8	2.6	21.2	10.6
66.1	5.8	24.2	8.6
214.7	24	4	17.4
23.8	35.1	65.9	9.2
97.5	7.6	7.2	9.7



QUESTIONS REGARDING ADVERTISING DATA

- Is there a relationship between advertising budget and sales?
  - If the evidence is weak, then one might argue that no money should be spent on advertising!
- How strong is the relationship between advertising budget and sales?
- Which media are associated with sales?
  - Are all three media - TV, radio, and newspaper - associated with sales, or are just one or two of the media associated?
- How large is the association between each medium and sales?
  - For every dollar spent on advertising in a particular medium, by what amount will sales increase?



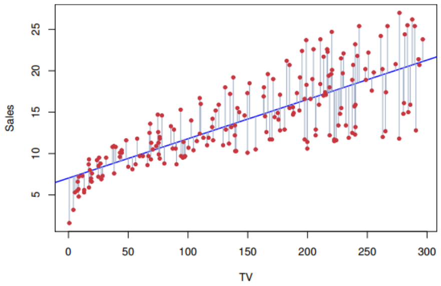
QUESTIONS REGARDING ADVERTISING DATA (CONT'D)

- How accurately can we predict future sales?
  - For any given level of television, radio, or newspaper advertising, what is our prediction for sales
- Is the relationship linear?
  - If there is approximately a straight-line relationship between advertising expenditure in the various media and sales, then linear regression is an appropriate tool.
- Is there synergy among the advertising media?
  - Perhaps spending $50,000 on television advertising and $50;000 on radio advertising is associated with higher sales than allocating $100,000 to either television or radio individually.
  - In marketing, this is known as a synergy effect, while in statistics it is called an interaction effect.

SIMPLE LINEAR REGRESSION

- Simple linear regression is a very straightforward approach for predicting a quantitative response Y on the basis of a single predictor variable X. It assumes that there is approximately a linear relationship between X and Y .
- sales ≈ βo + β1 × TV
- β0 and β1 are two unknown constants that represent the intercept and slope terms in the linear model. Together, β0 and β1 are known as the model coefficients or parameters.
- Once we have used our training data to produce estimates β^$_0$ and β^$_1$ for the model coefficients, we can predict future sales on the basis of a particular value of TV advertising.
- y^ = β^$_0$ + β^$_1$x
- The hat symbol, ^ , is used to denote the predicted value of the response.

SIMPLE LINEAR REGRESSION (CONT'D)



- Our goal is to obtain coefficient estimates β^0 an β^1 such that the linear model fits the available data well.
- In other words, we want to find an intercept β^0 and a slope β^1 such that the resulting line is as close as possible to the data points.
- There are a number of ways of measuring closeness. However, by far the most common approach involves minimizing the least squares criterion. Alternative approaches will be considered in the coming weeks (Chapter 6 of the ISL book).

ESTIMATING THE COEFFICIENTS

- We define the residual sum of squares (RSS) or sum of squared error (SSE)

$$\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2,$$

or equivalently as

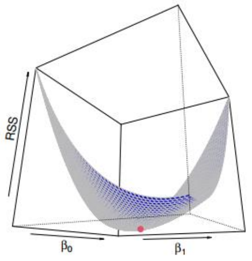
$$\mathrm{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \cdots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2.$$

- The least squares approach chooses β0 and β1 to minimize the RSS.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

ESTIMATING THE COEFFICIENTS (CONT'D)

- β_0 = 7.03 and β_1 = 0.0475
- According to this approximation, an additional $1,000 spent on TV advertising is associated with selling approximately 47.5 additional units of the product.
- What are the coefficients for Radio and Newspaper?



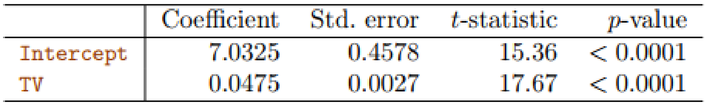
ASSESSING MODEL ACCURACY: RSE

$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}.$$

- The RSE (Residual Standard Error) is considered a measure of the *lack of fit* of a model to the data.
- In literature, this term is often referred to as "Root Mean Squared Error" (RMSE).
- If the predictions obtained using the model are very close to the true outcome values then RSE will be small, and we can conclude that the model fits the data very well.
- On the other hand, if ^*yi* is very far from *yi* for one or more observations, then the RSE may be quite large, indicating that the model doesn't fit the data well.

INTERPRETING P-VALUES

- H0 :There is no relationship between X and Y vs.
- Ha :There is some relationship between X and Y : (3.13)
- Mathematically,
- H0 : β1 = 0 versus Ha : β1 ≠ 0;
- if we see a small p-value, then we can infer that there is an association between the predictor and the response. We reject the null hypothesis -- that is, we declare a relationship to exist between X and Y -- if the p-value is small enough.



	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

ASSESSING MODEL ACCURACY: R2

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

where $\mathrm{TSS} = \sum (y_i - \bar{y})^2$ is the *total sum of squares*.

- The RSE provides an absolute measure of lack of fit of the model to the data. But since it is measured in the units of *Y* , it is not always clear what constitutes a good RSE.
- The *R*2 statistic provides an alternative measure of fit. It takes the form of a *proportion* - the proportion of variance explained - and so it always takes on a value between 0 and 1, and is independent of the scale of *Y*.
- TSS – RSS measures the amount of variability in the response that is explained (or removed) by performing the regression, and *R*2 measures the *proportion of variability in Y that can be explained using X*.

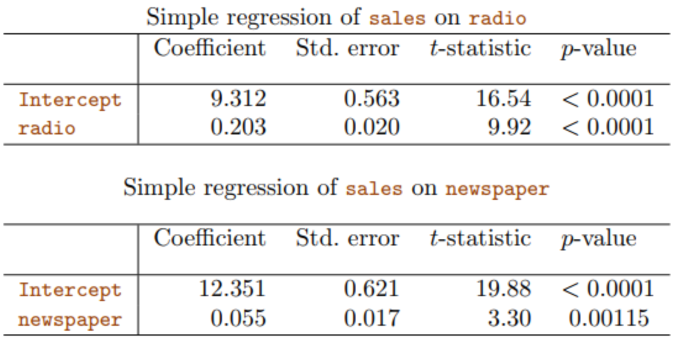
ADJUSTED R^2

- The challenge is to select a model that is not too simplistic in terms of excluding important parameters (the model is *under-fit*), nor overly complex thereby modeling random noise (the model is *over-fit*).
- Several criteria for evaluating and comparing models are based on metrics computed from the training data:
- One popular criterion is the *adjusted R*^{2}.

R^2 VS ADJUSTED R^2

- $R^2_{adj} = 1 – [(n – 1)/(n – p – 1)](1 – R^2)$
- Where n = number of records and p = number of variables
- Like R^2, higher values of adjusted R^2 indicate better fit.
- Unlike R^2, which does not account for the number of predictors used, adjusted R^2 uses a penalty on the number of predictors.
- This avoids the artificial increase in R^2 that can result from simply increasing the number of predictors but not the amount of information.

SIMPLE LINEAR REGRESSION (CONT'D)



Simple regression of sales on radio

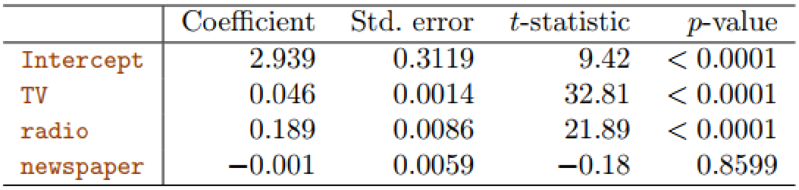
	Coefficient	Std. error	*t*-statistic	*p*-value
Intercept	9.312	0.563	16.54	< 0.0001
radio	0.203	0.020	9.92	< 0.0001

Simple regression of sales on newspaper

	Coefficient	Std. error	*t*-statistic	*p*-value
Intercept	12.351	0.621	19.88	< 0.0001
newspaper	0.055	0.017	3.30	0.00115

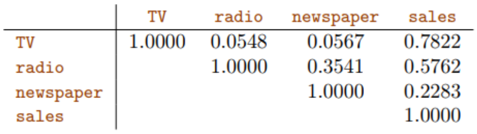
MULTIPLE LINEAR REGRESSION

- Unlike the simple linear regression estimates, the multiple regression coefficient estimates have somewhat complicated forms that are most easily represented using matrix algebra.
- For this reason, we do not provide them here.
- Any statistical software package can be used to compute these coefficient estimates.



	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	−0.001	0.0059	−0.18	0.8599

MULTIPLE LINEAR REGRESSION (CONT'D)



	TV	radio	newspaper	sales
TV	1.0000	0.0548	0.0567	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1.0000

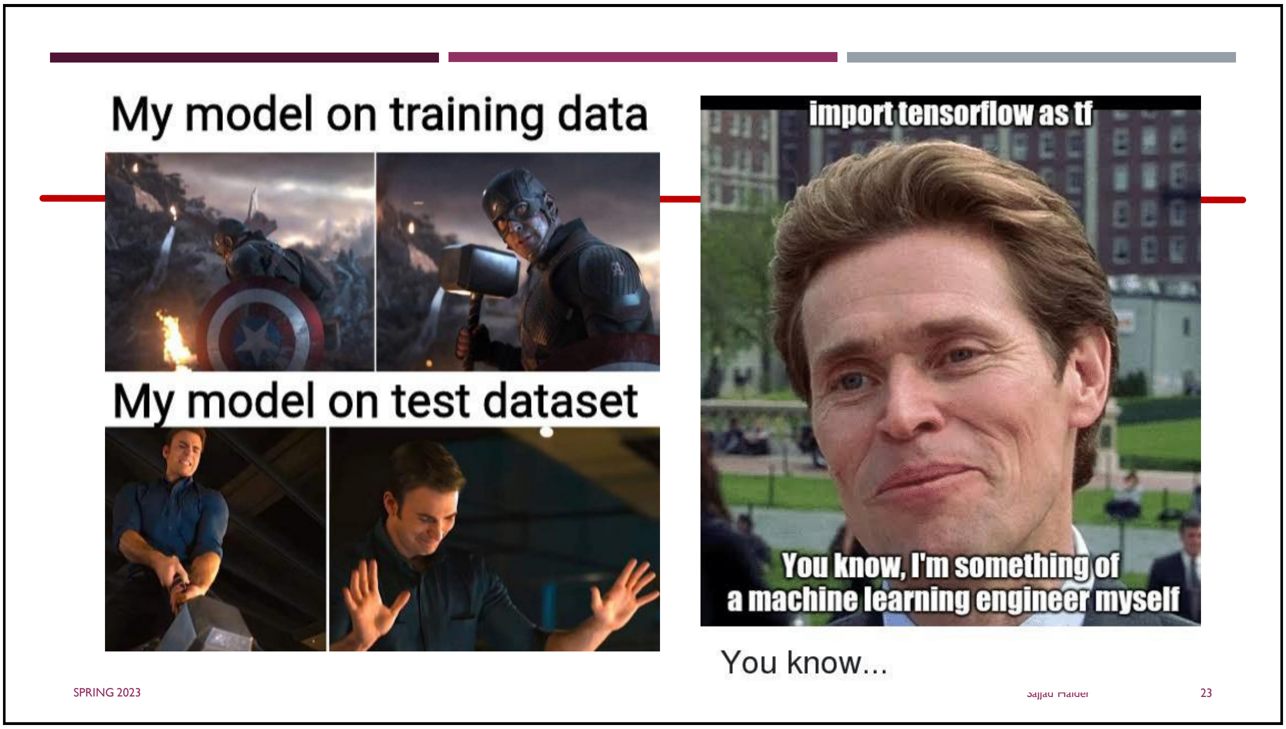
- Does it make sense for the multiple regression to suggest no relationship between sales and newspaper while the simple linear regression implies the opposite?
- Consider the correlation matrix for the three predictor variables and response variable. Notice that the correlation between radio and newspaper is 0:35.
- This indicates that markets with high newspaper advertising tend to also have high radio advertising.
- Now suppose that the multiple regression is correct and newspaper advertising is not associated with sales, but radio advertising is associated with sales.

MULTIPLE LINEAR REGRESSION (CONT'D)

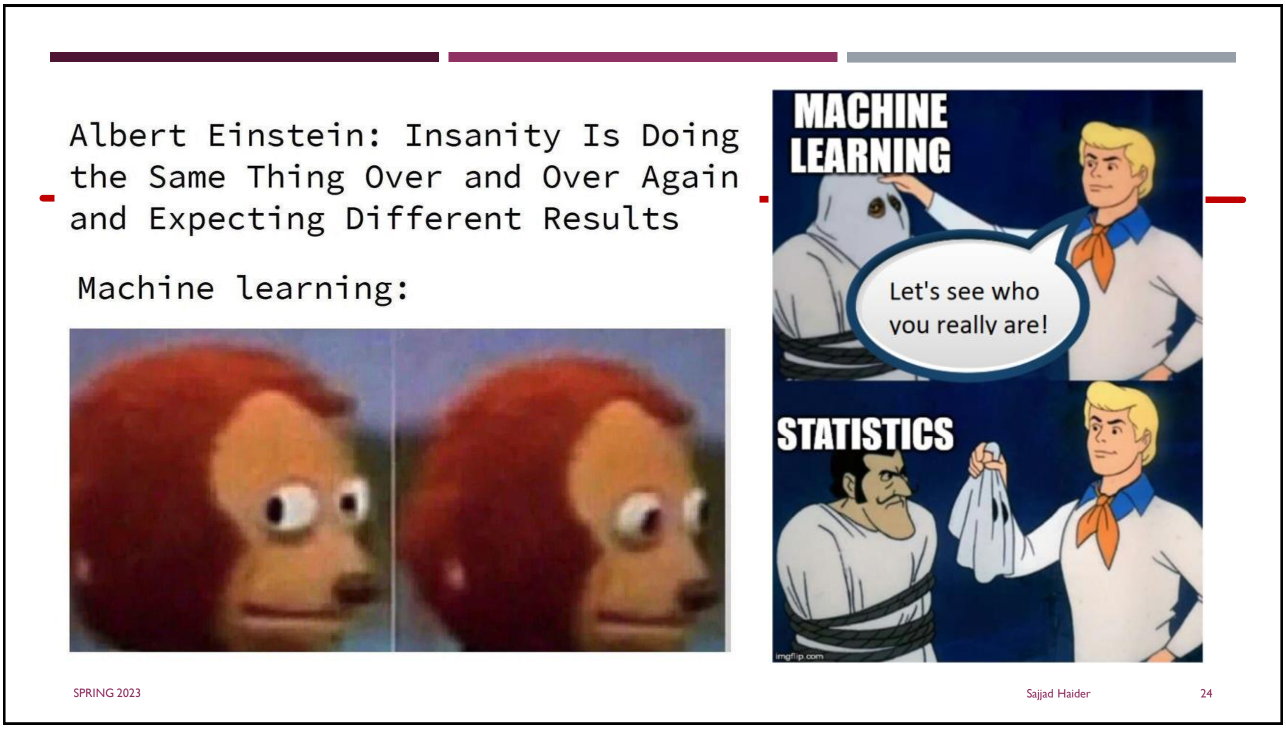
- Then in markets where we spend more on radio our sales will tend to be higher, and as our correlation matrix shows, we also tend to spend more on newspaper advertising in those same markets.
- Hence, in a simple linear regression which only examines sales versus newspaper, we will observe that higher values of newspaper tend to be associated with higher values of sales, even though newspaper advertising is not directly associated with sales.
- So newspaper advertising is a surrogate for radio advertising; newspaper gets "credit" for the association between radio on sales.

PYTHON DEMO

SINGLE AND MULTIPLE LINEAR REGRESSION USING SKLEARN AND STATSMODEL LIBRARIES


My model on training data
My model on test dataset
import tensorflow as tf
You know, I'm something of a machine learning engineer myself

You know...



Albert Einstein: Insanity Is Doing the Same Thing Over and Over Again and Expecting Different Results

Machine learning:

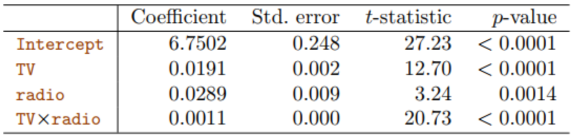
VARIABLE/FEATURE SELECTION

- The task of determining which predictors are associated with the response, in order to fit a single model involving only those predictors, is referred to as variable selection.
- Unfortunately, there are a total of 2^P models that contain subsets of p variables. This means that even for moderate p, trying out every possible subset of the predictors is infeasible.
- if $p = 30$, then we must consider $2^{30} = 1{,}073{,}741{,}824$ models!
- Forward selection. We begin with the *null model* - a model that contains an intercept but no predictors. We then fit p simple linear regressions and add to the null model the variable that results in the lowest RSS. We then add to that model the variable that results in the lowest RSS for the new two-variable model. This approach is continued until some stopping rule is satisfied.

VARIABLE/FEATURE SELECTION (CONT'D)

- Backward selection. We start with all variables in the model, and remove the variable with the largest p-value. The new (p - 1)-variable model is fit, and the variable with the largest p-value is removed. This procedure continues until a stopping rule is reached.
- Mixed selection (Step-wise Regression). This is a combination of forward and backward selection. We start with no variables in the model, and as with forward selection, we add the variable that provides the best fit. We continue to add variables one-by-one. If at any point the p-value for one of the variables in the model rises above a certain threshold, then we remove that variable from the model. We continue to perform these forward and backward steps until all variables in the model have a sufficiently low p-value, and all variables outside the model would have a large p-value if added to the model.

INTERACTION MODEL



	Coefficient	Std. error	t-statistic	p-value
Intercept	6.7502	0.248	27.23	< 0.0001
TV	0.0191	0.002	12.70	< 0.0001
radio	0.0289	0.009	3.24	0.0014
TV×radio	0.0011	0.000	20.73	< 0.0001

- Suppose that spending money on radio advertising actually increases the effectiveness of TV advertising, so that the slope term for TV should increase as radio increases.
- In this situation, given a fixed budget of $100;000, spending half on radio and half on TV may increase sales more than allocating the entire amount to either TV or to radio.
- sales = $\beta_0 + \beta_1$ × TV + β_2 × radio + β_3 × (radio × TV) + ε
- If the interaction between $X1$ and $X2$ seems important, then we should include both $X1$ and $X2$ in the model even if their coefficient estimates have large p-values. The rationale for this principle is that if $X1 \times X2$ is related to the response, then whether or not the coefficients of $X1$ or $X2$ are exactly zero is of little interest.

PYTHON DEMO

POLYNOMIAL AND INTERACTION MODEL, FEATURE SELECTION