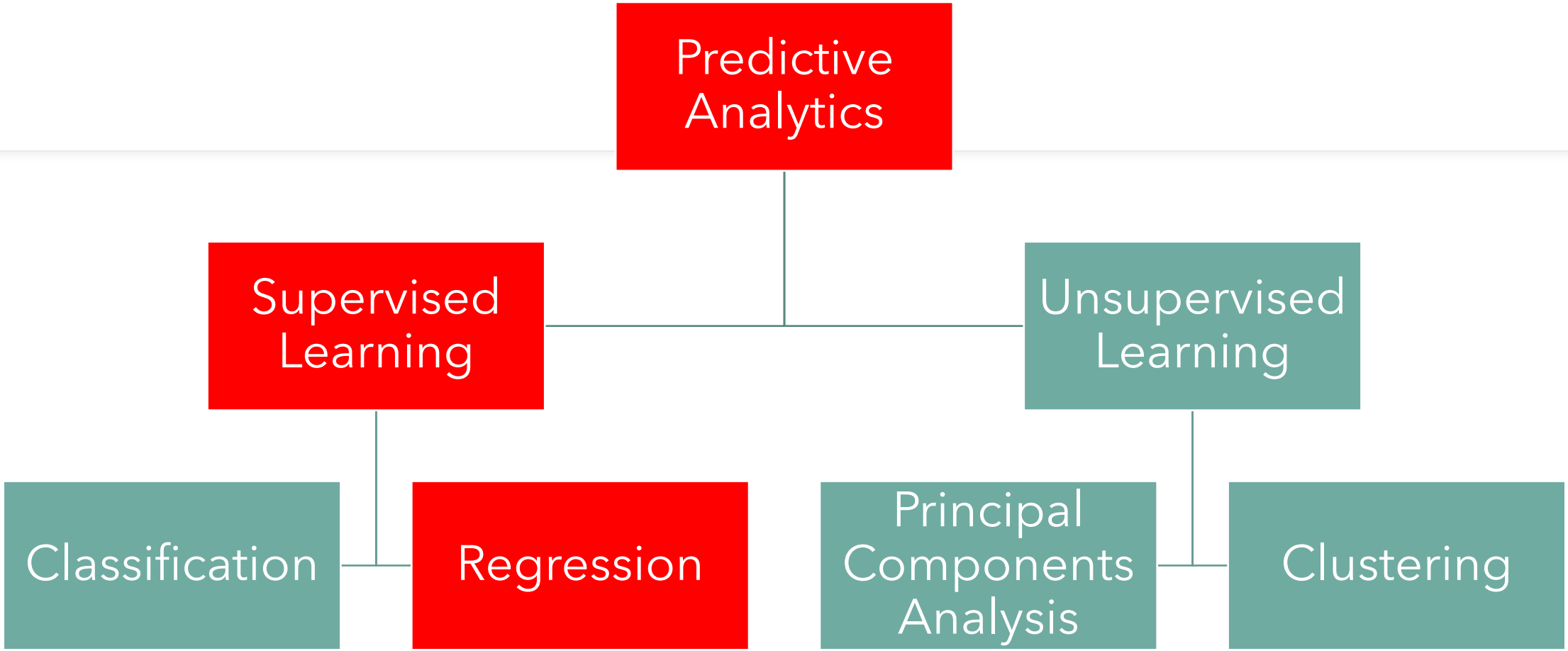
Linear Regression Review

Today's Lecture: Key Concepts

1. Review of linear regression model and how it works
2. Know how to answer key questions for linear regression in relation to business problem
   - All statistical tests
   - Know how linear regression (ordinary least squares) is solved
3. Interpret results and model performance


Predictive Analytics
Supervised Learning
Unsupervised Learning
Classification
Regression
Principal Components Analysis
Clustering

You've been asked to suggest a marketing strategy that will boost future sales.

What information would be useful to provide a recommendation for a marketing plan?

Is there a relationship between advertising budget and sales?

Which media contribute to sales?

How well does the model predict new data?

Is there a relationship between advertising budget (marketing strategy) and sales?

- Does the data provide evidence of an association between advertising and sales?

- If the evidence is weak, then one could make the argument that no money should be spent on advertising!

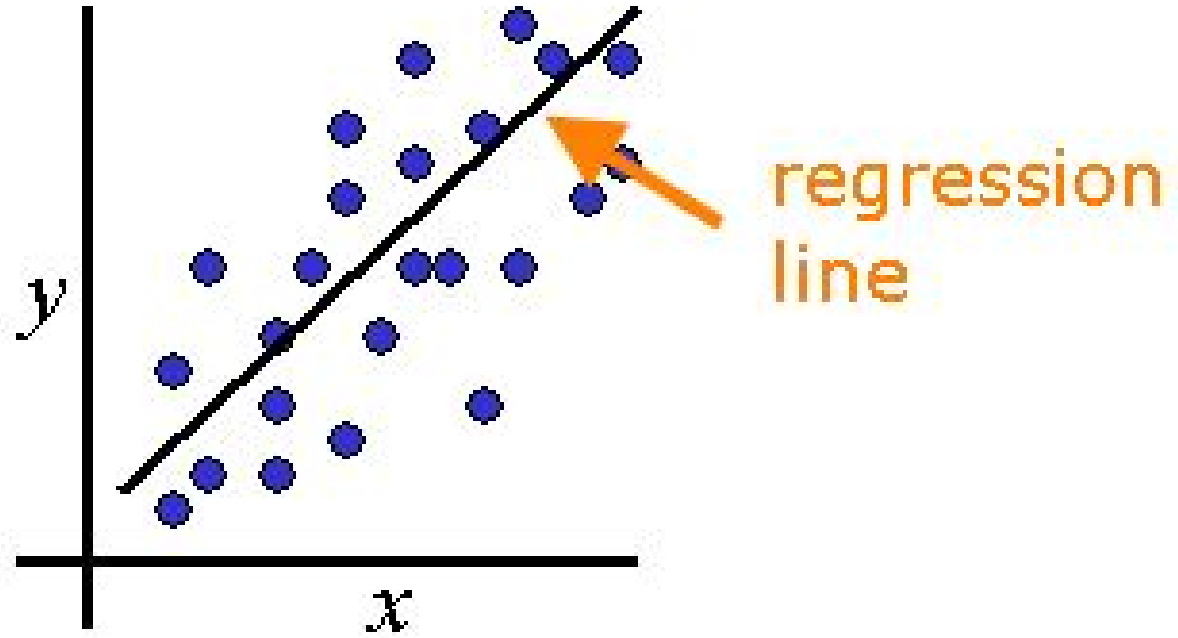
Which media contribute to sales?

- Do all three media contribute to sales—newspaper, radio, and TV or does just one or two contribute?

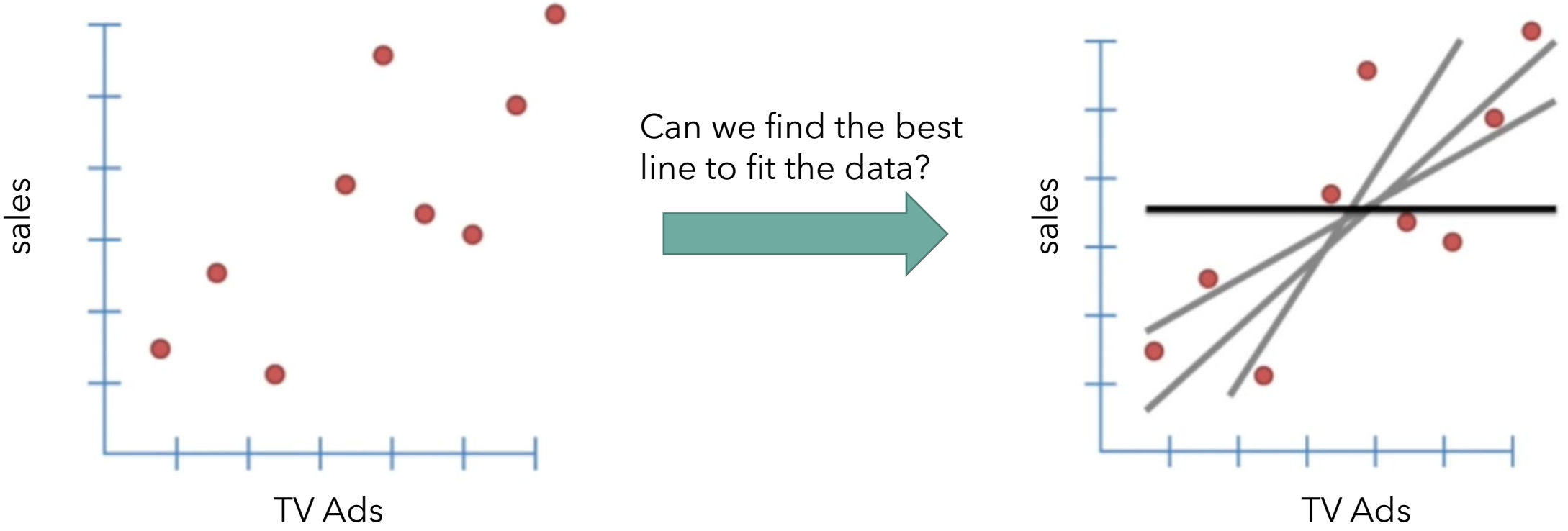
How well does the model predict new data?

- Does our model generalize well?
- i.e. how accurately can we predict sales for new data ?

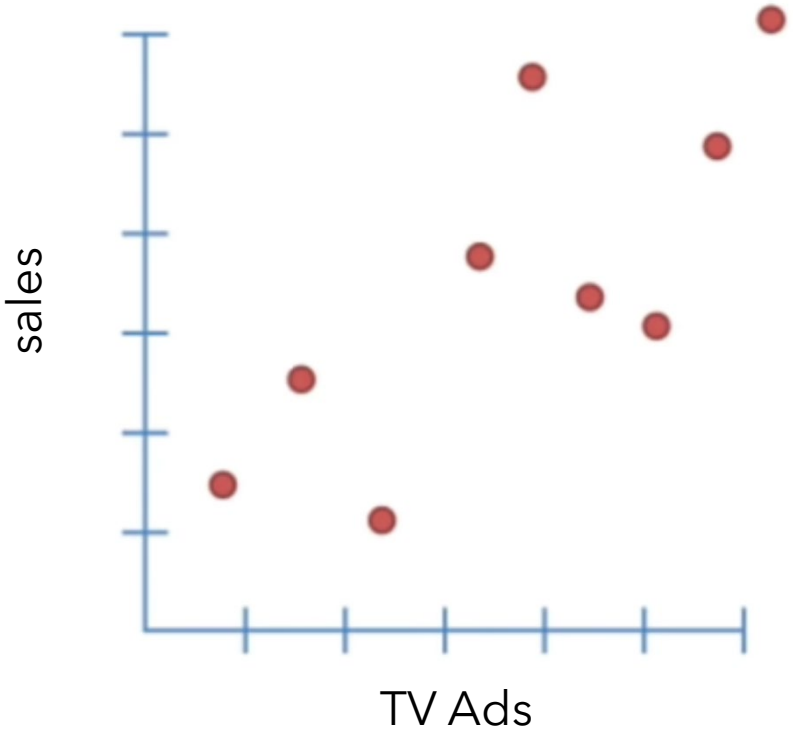
Using data on past sales performance of several advertising channels, **linear regression can be used to answer each of these key questions**!



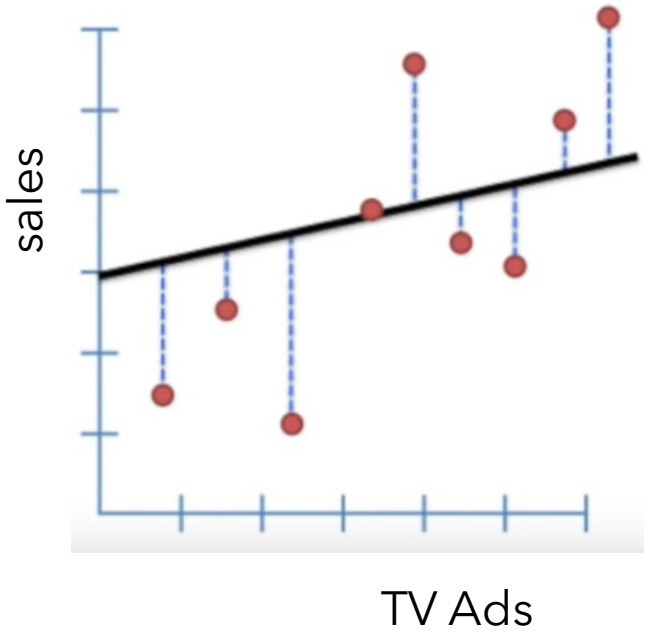
Linear Regression Graphically



Linear Regression Graphically



Can we find the best line **to minimize the sum of squared errors (errors=length of the blue lines between points the line)**?

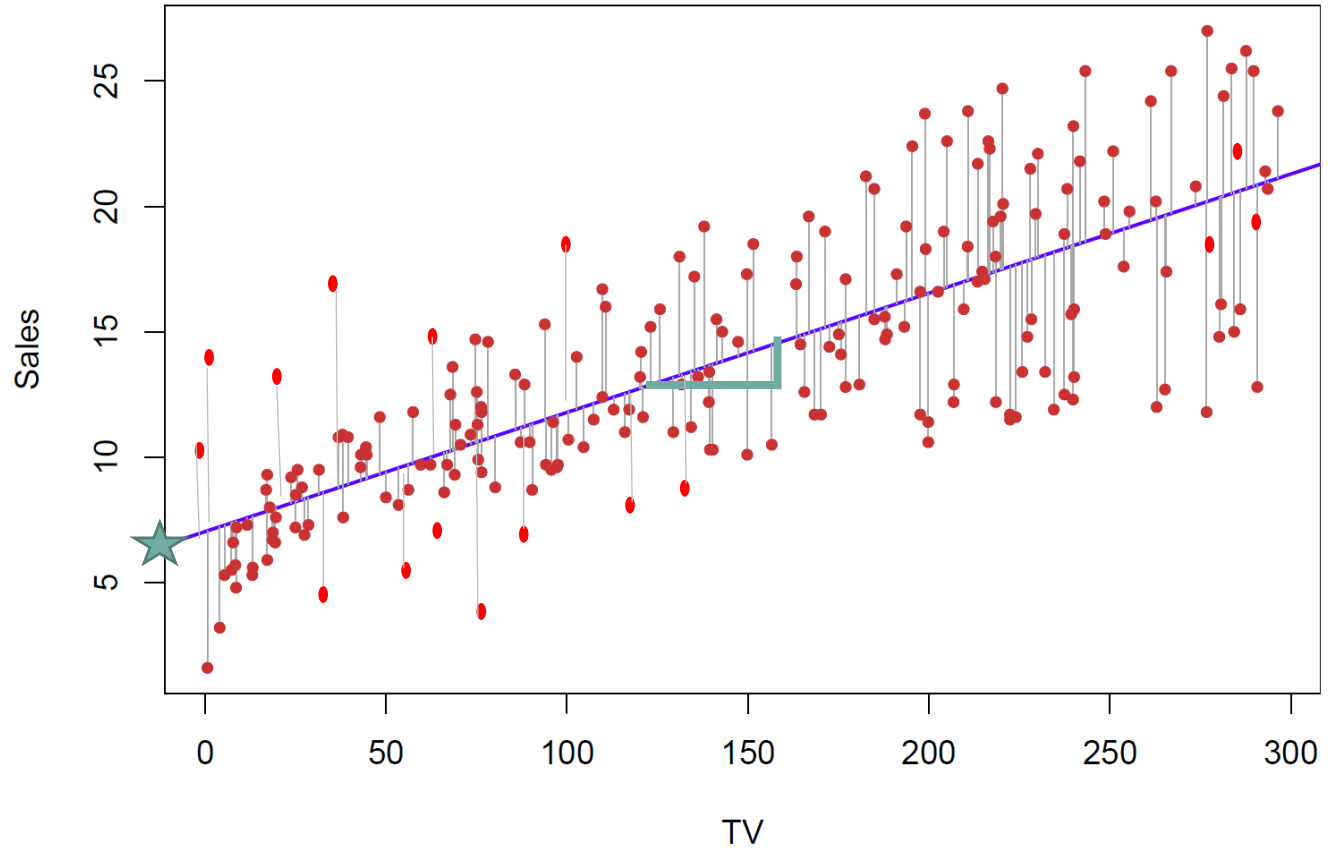


The Linear Regression Model from Population

$$Y = \beta_0 + \beta_1 X_1 + \epsilon$$

- Y is response variable (sales)
- X_1 is the predictor variable (TV ads)
- β_0 is population parameter for y-intercept
- β_1 is population parameter for slope
- ϵ is the error term or distance between the actual value and the regression equation (ϵ is normal random variable with mean 0 and variance σ^2)

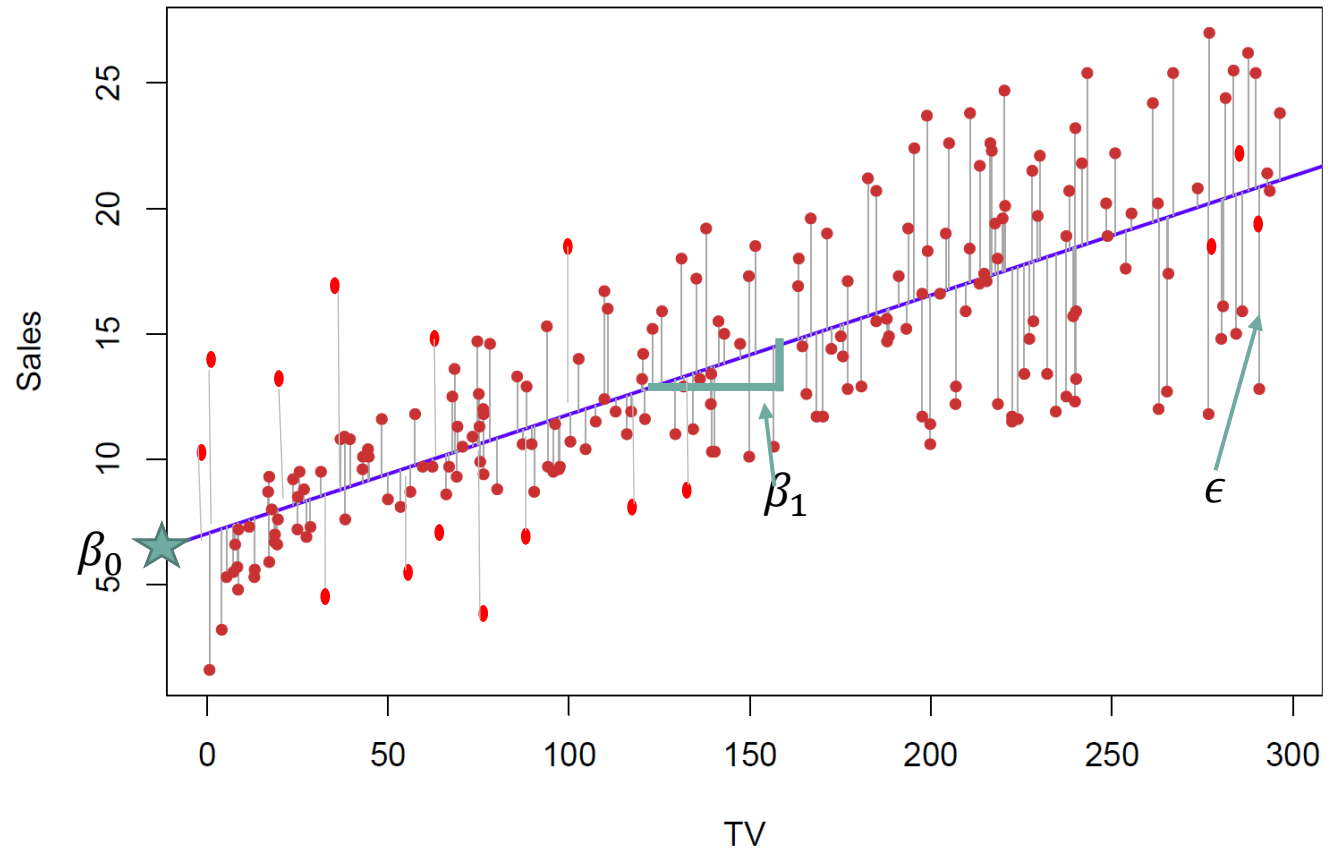
The Linear Regression Model from Population



True population line
$Y = \beta_0 + \beta_1 X_1$

What is β_0?
What is β_1?
What is ϵ?

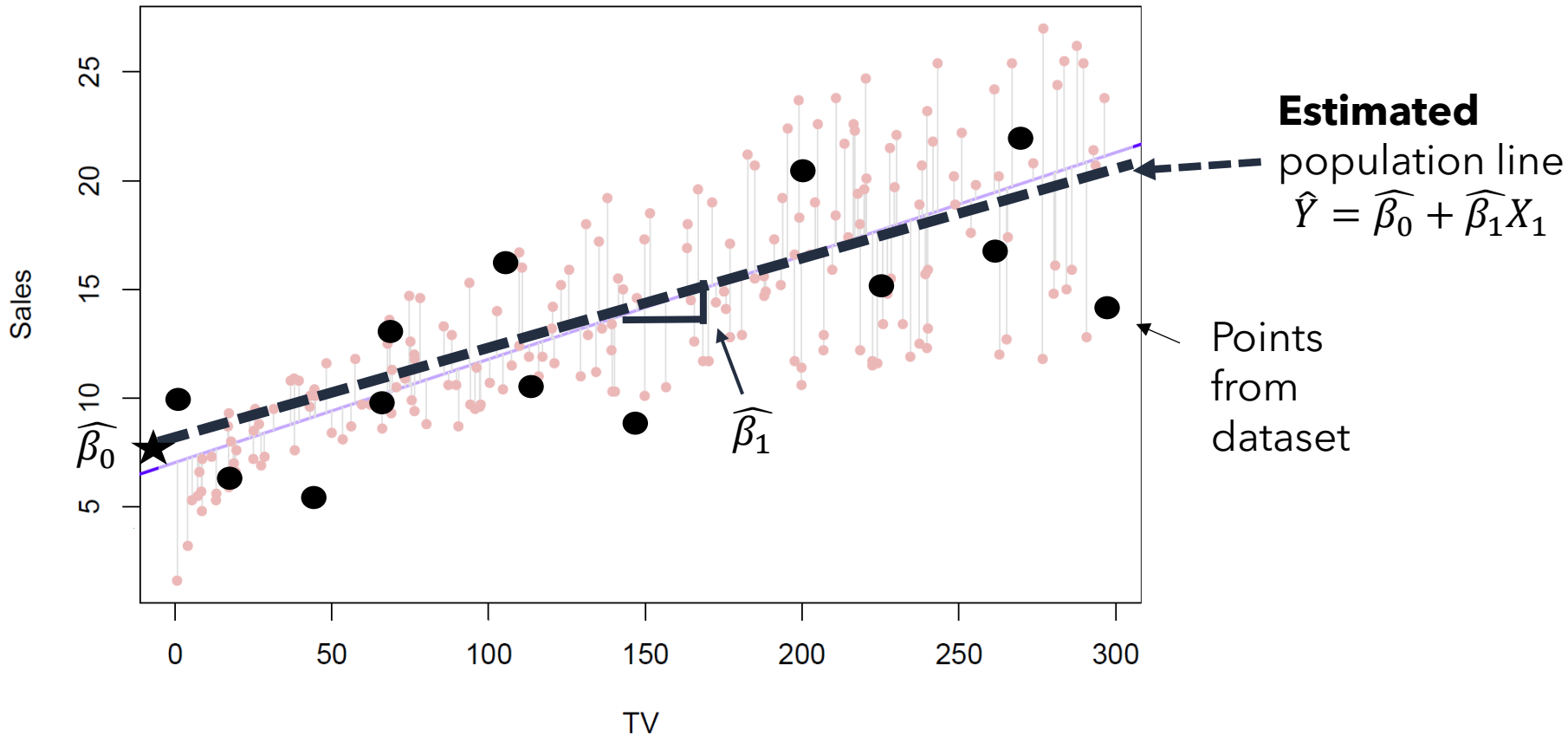
The Linear Regression Model from Population



True population line
$Y = \beta_0 + \beta_1 X_1$

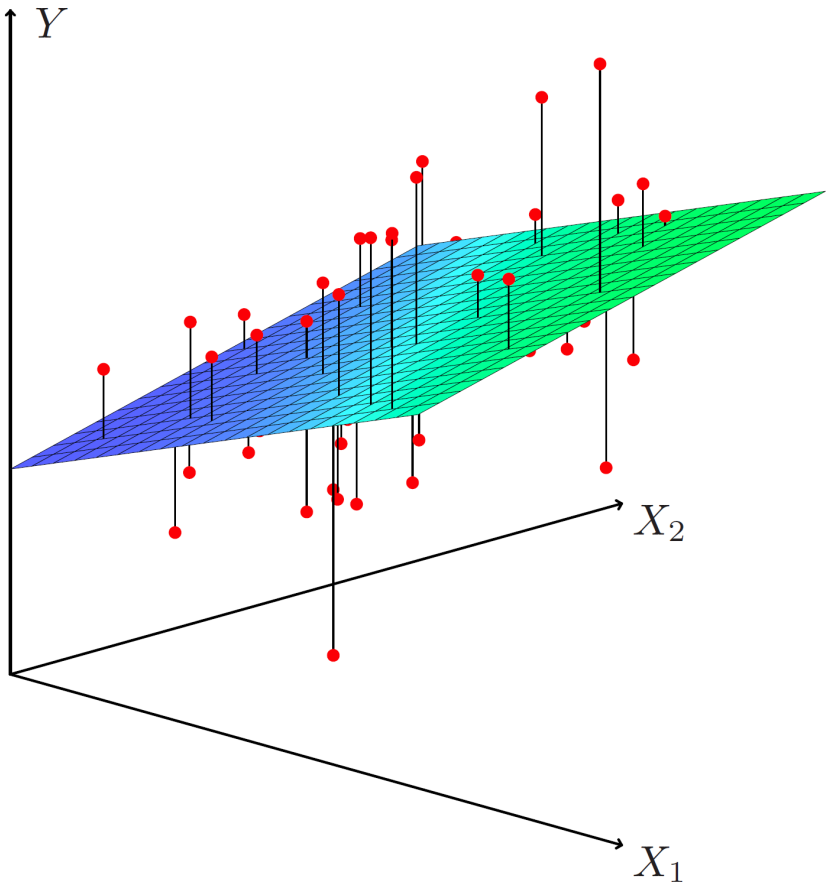
What is β_0?
What is β_1?
What is ϵ?

Our goal is to estimate this population line from a dataset.



With 2 predictor variables, can we find the best plane to fit the data?

In other words, can we find the best plane **to minimize the squared black lines from the points to the plane**?



(Multiple) Linear Regression Model

- Linear regression model for **population**:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

  - Y is response variable
  - $X_1, X_2, \ldots, X_p$ are predictor variables
  - $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ are population parameters
  - ϵ is the error term. ϵ is random variable with mean 0 and variance σ^2

- Goal: Estimated multiple linear regression model from **dataset**:

$$\hat{Y} = \widehat{\beta_0} + \widehat{\beta_1} X_1 + \widehat{\beta_2} X_2 + \cdots + \widehat{\beta_p} X_p$$

  - $\hat{Y}$ is predicted response value
  - $\widehat{\beta_0}, \widehat{\beta_1}, \ldots, \widehat{\beta_p}$ are coefficients estimates or estimated values of parameter

The Linear Regression Model

$X_1, \ldots, X_p$ are the predictor variables.

$$\hat{Y} = \widehat{\beta_0} + \widehat{\beta_1} X_1 + \widehat{\beta_2} X_2 + \cdots + \widehat{\beta_p} X_p$$

- β_j is the estimated average increase/decrease on Y of a one unit increase in X_j holding all other variables constant.
- β_o is the y-intercept or the estimated average value for Y if all the X's are zero.
- $\hat{Y}$ is the predicted response variable from the model

How do we find best fit line?

We estimate $\beta_0, \beta_1, \cdots, \beta_j$ by finding the regression model that minimizes the sum of squared errors on the training set.

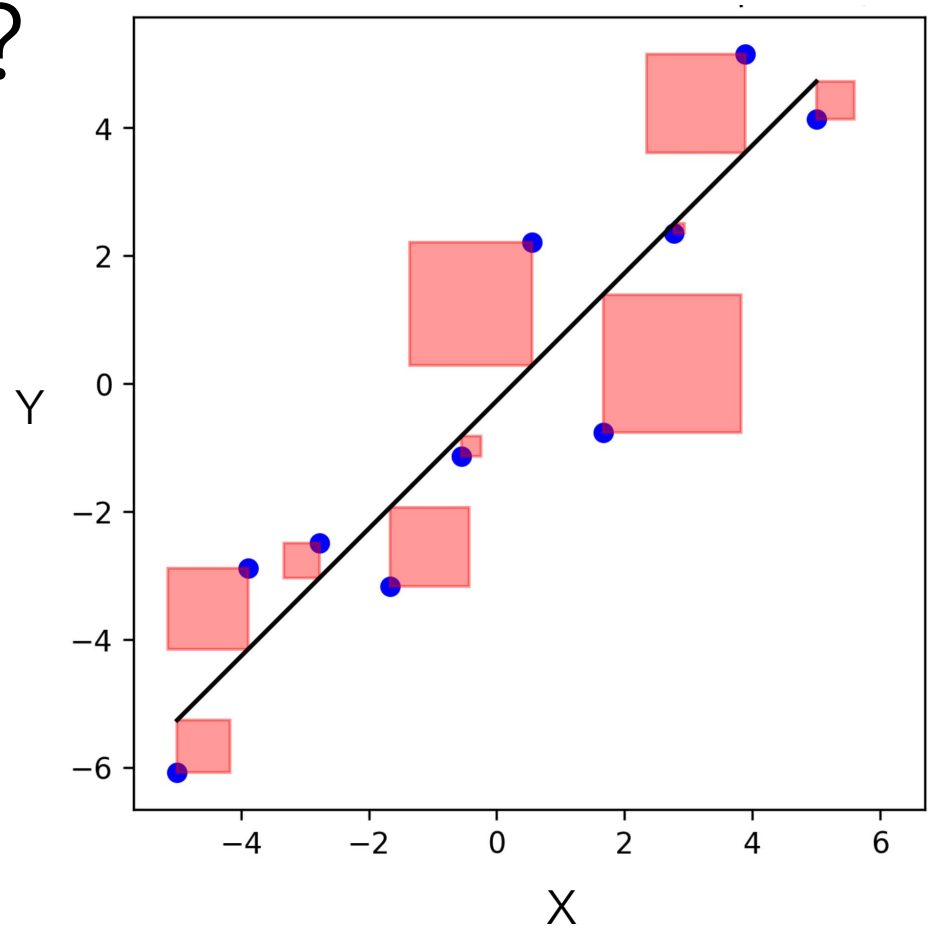
Find best regression model to minimize

$$\sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2$$

Sum over all data points

Response data point

Predicted response from regression line



* Everything in pink are the sum of squared errors. We find the best line to minimize the pink area.

Back to the client's problem…

What is a marketing plan for next year that will result in high product sales?

Key Questions for Linear Regression Problems:

What information would be useful to provide a recommendation for a marketing plan?

Is there a relationship between advertising budget and sales?

Which media contribute to sales?

How well does the model predict new data?

Exercise: Using **python**, run linear regression(ordinary least squares) to predict sales on Advertising.csv in lab folder.

Is there a relationship between advertising budget and sales?

In linear regression model, is at least one predictor variable useful in predicting the response?

To answer this question, we use the F-test.

Null Hypothesis: All coefficients $\beta_1 = \beta_2 = \cdots = \beta_p$ are equal to 0.

*p is the total number of predictor variables.

Why 0?

Analysis of Variance F-test

- Null Hypothesis: All coefficients $\beta_1 = \beta_2 = \cdots = \beta_p$ are equal to 0.
- Alternative Hypothesis: ???
- What does this mean in nontechnical language?

Analysis of Variance F-test

- Null Hypothesis: H_0: All coefficients $\beta_1 = \beta_1 = \cdots = \beta_p$ are equal to 0.
- Alternative Hypothesis: H_a: at least one $\beta_j \neq 0$
- What does this mean in nontechnical language?
  - Null Hypothesis: H_0: All coefficients $\beta_1 = \beta_1 = \cdots = \beta_p$ are equal to 0. → No predictor variables are related to the response
  - Alternative Hypothesis: H_a: at least one $\beta_j \neq 0$ → at least one predictor variable is related to response.

Analysis of Variance F-test

Another way of saying predictor variables are related to the response: the predictor variables (regression) explain a lot of variability in the response variable

Total variation in response =

variation explained by regression + unexplained remaining variation

How can we determine if more of the total variation is **explained by regression** than the **unexplained remaining variation**?

Analysis of Variance F-test

Total variation in response = variation explained by regression + unexplained remaining variation

The hypothesis test is performed using the *F-statistic,*

$$F = \frac{\text{Estimate of variance captured by regression model}}{\text{Estimate of variance that is unexplained}}$$

Does a larger or smaller F determine if more of the total variation is explained by regression?

Analysis of Variance F-test

Total variation in response = variation explained by regression + unexplained remaining variation

The hypothesis test is performed using the *F-statistic,*

$$F = \frac{\text{Estimate of variance captured by regression model}}{\text{Estimate of variance that is unexplained}}$$

The larger the F statistic the more likely to reject null hypothesis.
What is the F statistic telling us in the advertising data?

Is there a relationship between advertising budget and sales?

OLS Regression Results

Dep. Variable:	sales	R-squared:	0.891
Model:	OLS	Adj. R-squared:	0.889
Method:	Least Squares	F-statistic:	423.4
Date:		Prob (F-statistic):	1.02e-74
Time:		Log-Likelihood:	-314.14
No. Observations:	160	AIC:	636.3
Df Residuals:	156	BIC:	648.6
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.9045	0.369	7.877	0.000	2.176	3.633
TV	0.0455	0.002	28.499	0.000	0.042	0.049
radio	0.1882	0.010	18.941	0.000	0.169	0.208
newspaper	0.0008	0.007	0.114	0.909	-0.012	0.014

- In linear regression model, is at least one predictor variable useful in predicting the response?

1. H_0: All coefficients $\beta_1 = \beta_1 = \cdots = \beta_p$ are equal to 0. H_a: at least one $\beta_j \neq 0$.
2. The hypothesis test is performed using the *F-statistic*
3. Assuming null is true, the probability of observing any number equal to F value or larger (known as the ??).
   - If ?? is less than 0.05, at least one predictor variable is related to the response

Is there a relationship between advertising budget and sales?

OLS Regression Results

Dep. Variable:	sales	R-squared:	0.891
Model:	OLS	Adj. R-squared:	0.889
Method:	Least Squares	F-statistic:	423.4
Date:		Prob (F-statistic):	1.02e-74
Time:		Log-Likelihood:	-314.14
No. Observations:	160	AIC:	636.3
Df Residuals:	156	BIC:	648.6
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.9045	0.369	7.877	0.000	2.176	3.633
TV	0.0455	0.002	28.499	0.000	0.042	0.049
radio	0.1882	0.010	18.941	0.000	0.169	0.208
newspaper	0.0008	0.007	0.114	0.909	-0.012	0.014

- In linear regression model, is at least one predictor variable useful in predicting the response?

1. H_0: All coefficients $\beta_1 = \beta_1 = \cdots = \beta_p$ are equal to 0. H_a: at least one $\beta_j \neq 0$.
2. The hypothesis test is performed using the *F-statistic*
3. Assuming null is true, the probability of observing any number equal to F value or larger (known as the ??).
   - If p-value is less than 0.05, at least one predictor variable is related to the response

Which media contribute to sales?

- If we reject the null-hypothesis in the F test, then we use the t-test hypothesis test to determine which media contribute to sales
- H_0: $\beta_j = 0$
- H_a: $\beta_j \neq 0$

- Why 0?

Hypothesis test to determine if a specific predictor variable is related to the response: t - test

- To test the null hypothesis, we need to determine if β_j is sufficiently far from 0 so that we can be confident that β_j is not 0. How far is far enough?
- This depends on the estimated standard deviation of the coefficient estimate $\widehat{\beta_j}$, known as $SE(\beta_j)$.
  - The standard error of an estimator reflects how it varies with different datasets.

Hypothesis test to determine if a specific predictor variable is related to the response

1. $H_0: \beta_j = 0$

   $H_a: \beta_j \neq 0$

2. We compute the t statistic $t = \frac{\widehat{\beta_j} - 0}{SE(\widehat{\beta_j})}$

   This t statistic measures the number of standard deviations that $\widehat{\beta_j}$ is away from 0. Larger t more likely to be different from 0.

3. Compute the probability of observing any number equal to |t| or larger in absolute value (known as the p-value). A small probability indicates that it is unlikely to observe such a strong relationship due to chance.

   What are the t-statistics telling us in advertising data?

Which media contribute to sales?

OLS Regression Results

Dep. Variable:	sales	R-squared:	0.891
Model:	OLS	Adj. R-squared:	0.889
Method:	Least Squares	F-statistic:	423.4
Date:		Prob (F-statistic):	1.02e-74
Time:		Log-Likelihood:	-314.14
No. Observations:	160	AIC:	636.3
Df Residuals:	156	BIC:	648.6
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.9045	0.369	7.877	0.000	2.176	3.633
TV	0.0455	0.002	28.499	0.000	0.042	0.049
radio	0.1882	0.010	18.941	0.000	0.169	0.208
newspaper	0.0008	0.007	0.114	0.909	-0.012	0.014

- In linear regression model, is a specific predictor variable related to the response?

1. $H_0: \beta_j = 0. H_a: \beta_j \neq 0.$
2. The hypothesis test is performed using the *t-statistic*
3. Compute the probability of observing any number equal to ltl value or larger is (known as the p-value).

- If p-value is less than 0.05, conclude the predictor variable X_j is related to the response.

Which media contribute to sales?

OLS Regression Results

Dep. Variable:	sales	R-squared:	0.891
Model:	OLS	Adj. R-squared:	0.889
Method:	Least Squares	F-statistic:	423.4
Date:		Prob (F-statistic):	1.02e-74
Time:		Log-Likelihood:	-314.14
No. Observations:	160	AIC:	636.3
Df Residuals:	156	BIC:	648.6
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.9045	0.369	7.877	0.000	2.176	3.633
TV	0.0455	0.002	28.499	0.000	0.042	0.049
radio	0.1882	0.010	18.941	0.000	0.169	0.208
newspaper	0.0008	0.007	0.114	0.909	-0.012	0.014

- Each coefficient is the expected change in sales (in thousands of units) for a +$1,000 change in that media budget, holding the other media budgets fixed.
- Intercept = 2.94: Predicted sales when TV = Radio = Newspaper = 0
- TV = +$1k TV → +0.046k sales (≈ +46 units)
- Radio +$1k radio → +0.189k sales (≈ +189 units)

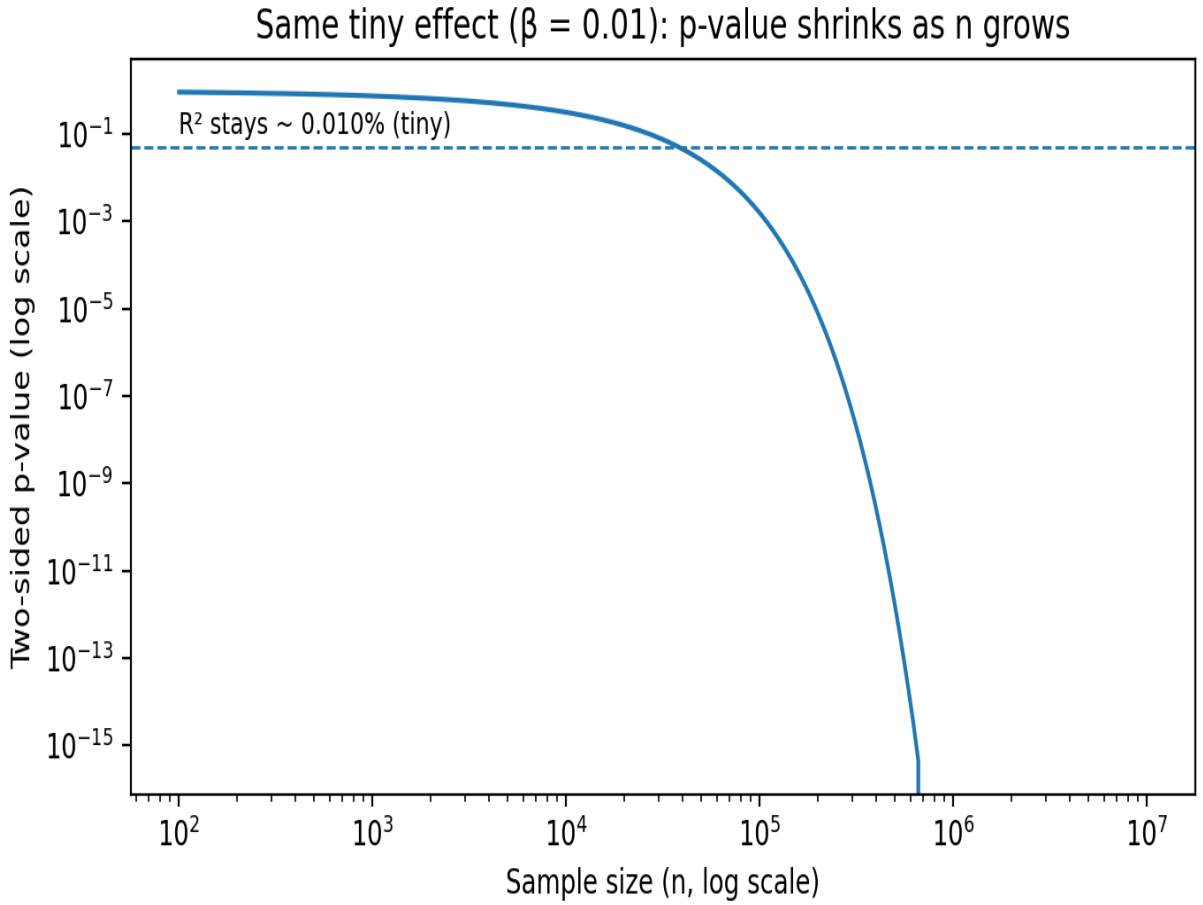
Units: sales are in 1,000 units; budgets are in $1,000s.

Big data can make tiny effects look “significant”

- The t-statistic is estimate ÷ standard error (SE)
- With bigger n, SE gets smaller (since SE denominator is 1/√n)
- So a practically tiny effect can become statistically significant when data get very large
- **<u>But prediction may barely improve so we need to check test set MSE</u>**

Example: $Y = 0.01 \cdot X + \text{noise}$ ($\sigma = 1$)

n	|t|	p-value	R^2
1,000	0.32	0.75	0.01%
1,000,000	10.0	1.5e−23	0.01%



Takeaway: statistical significance ≠ practical importance (or predicting well for new data).

How well does the model **predict new data**?

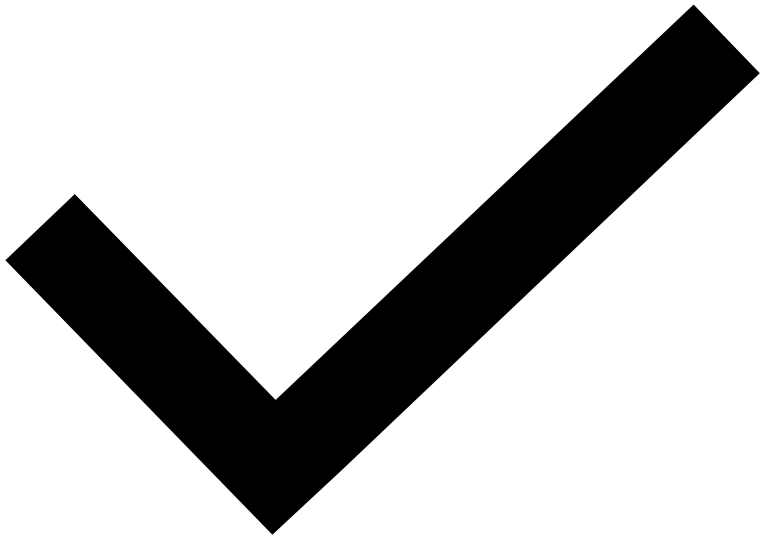
- Remember we are interested in predicting sales for observations **we haven't seen before!**

- To reduce overfitting, when a model learns the training data too well and thus struggles on unseen data, we divide the data into training and test sets.

How well does the model **predict new data**?

- How do we evaluate our model?
  - We build the linear regression model on the training set.
- **Then calculate mean squared error, MSE, on the testing set**.

Why MSE on testing set?



Consulting Question:

Given your demonstrated fluency with the fundamentals of data science, you are called in to help a client understand data science concepts.

How do we correctly assess the model on new predictions?

Which part of the data do we fit the model on, training or testing?

Training Data

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
204.1	32.9	46	19
195.4	47.7	52.9	22.4
67.8	36.6	114	12.5
281.4	39.6	55.8	24.4
69.2	20.5	18.3	11.3
147.3	23.9	19.1	14.6

Testing Data

Model: Sales = $\widehat{\beta_0}+\widehat{\beta_1} *$TV $+\widehat{\beta_2}*$Radio

What is the fitted model on the training data?

Fitted Model: Sales = ? + ? * TV + ? * Radio

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3.104656	0.315371	9.84	<.0001
TV	1	0.044579	0.001489	29.94	<.0001
radio	1	0.186902	0.008497	22.00	<.0001

How to we get Mean Squared Error (MSE) on the test set?

	TV	radio	newspaper	sales
Training Data	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
	204.1	32.9	46	19
Testing Data	195.4	47.7	52.9	22.4
	67.8	36.6	114	12.5
	281.4	39.6	55.8	24.4
	69.2	20.5	18.3	11.3
	147.3	23.9	19.1	14.6

Fitted Model: Sales = 3.105 +0.045 * TV + 0.187 * Radio

First, how do we get the prediction for Sales on the test set? What data do we plug into the fitted model?

How to we get Mean Squared Error (MSE) on the test set?

	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
Training Data	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
	204.1	32.9	46	19
	195.4	47.7	52.9	22.4
	67.8	36.6	114	12.5
Testing Data	281.4	39.6	55.8	24.4
	69.2	20.5	18.3	11.3
	147.3	23.9	19.1	14.6

What is the squared error for one test observation in red box?

sales prediction for test observation =
3.105 +0.045 * TV + 0.187 * Radio = ??

squared error = (sales prediction for test data – sales in test data)^2= ??

How to we get Mean Squared Error (MSE) on the test set?

	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
Training Data	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
	204.1	32.9	46	19
	195.4	47.7	52.9	22.4
	67.8	36.6	114	12.5
Testing Data	281.4	39.6	55.8	24.4
	69.2	20.5	18.3	11.3
	147.3	23.9	19.1	14.6

What is the squared error for one test observation in red box?

sales prediction for test data =
3.105 +0.045 * TV + 0.187 * Radio =
3.105 +0.045 * 67.8 + 0.187 * 36.6 = 13.00

Squared error =
(sales prediction for test data – sales in test data)^2 =
(13.00-12.5)^2 = 0.25

How to we get Mean Squared Error (MSE) on the test set?

	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
Training Data	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
	204.1	32.9	46	19
	195.4	47.7	52.9	22.4
	67.8	36.6	114	12.5
Testing Data	281.4	39.6	55.8	24.4
	69.2	20.5	18.3	11.3
	147.3	23.9	19.1	14.6

- Squared Error for one observation = (sales prediction for test data – sales from test data) ^ 2.
- Then average over all observations.

What is the Mean Squared Error on this test set?

Fitted Model from training data: Sales = 3.105 +0.045 * TV + 0.187 * Radio

Testing Data

TV	radio	newspaper	sales
281.4	39.6	55.8	24.4
69.2	20.5	18.3	11.3
147.3	23.9	19.1	14.6

Row 1: squared error = (sales prediction for test data – sales in test data)^2 = ??

Row 2: squared error = (sales prediction for test data – sales in test data)^2 = ??

Row 3: squared error = (sales prediction for test data – sales in test data)^2 = ??

Mean squared error (Test) = ??

What is the Mean Squared Error on this test set?

Fitted Model from training data: Sales = 3.105 +0.045 * TV + 0.187 * Radio

Testing Data

TV	radio	newspaper	sales
281.4	39.6	55.8	24.4
69.2	20.5	18.3	11.3
147.3	23.9	19.1	14.6

Row 1: squared error = (sales prediction for test data – sales in test data)^2 = 1.51

Row 2: squared error = (sales prediction for test data – sales in test data)^2 = 1.56

Row 3: squared error = (sales prediction for test data – sales in test data)^2 = 0.16

Mean squared error (Test) = 1.08

Dummy Variables for Regression

- Can we use words from categorical predictor variables in regression as numerical variables?
- For example: can we use “Urban” and “Not Urban” (category listings) in a regression equation?

Dummy Variables for Regression

- We must convert from categorical predictor variables to numerical predictor variables
- We simply create one or more indicator/dummy variable(s) that can be 0 or 1
  - The number of variables created depends on the number of categories of categorical variable.

Categorial Predictor Variables

- Can we use “Urban” and “Not Urban” (category listings) in a regression equation?
- Code them as indicator variables (dummy variables)
- For example, we can “code” Urban=0 and Not Urban= 1.

Interpretation

- Suppose we want to include income and urban to predict sales of a product.
- Then the regression equation is

$$Y_i \approx \beta_0 + \beta_1 Income_i + \beta_2 Urban$$

- β_2 is the average extra sales that <u>not urban</u> locations have for a given income level, since not urban is coded as 1.
  - Urban locations are the "baseline" since urban is coded as 0.

- Predictions

If NOT urban → $Predicted\ sales = \widehat{\beta_0} + \widehat{\beta_1} Income_i + \widehat{\beta_2} * 1$

If urban → $Predicted\ sales = \widehat{\beta_0} + \widehat{\beta_1} Income_i + \widehat{\beta_2} * 0$

Interpretation

- For Dummy Variables with the regression equation:

  $ResponseVariable \approx \beta_0 + \beta_1 X_1 + \beta_2 DummyVariable$

- The interpretation is:

  β_2 is the average extra *response variable* that *dummy variable = 1* has for a given level of other variables.

- Predictions

If *dummy variable = 1* → $Predicted\ response = \widehat{\beta_0} + \widehat{\beta_1} X_1 + \widehat{\beta_2}$

If *dummy variable = 0* → $Predicted\ response = \widehat{\beta_0} + \widehat{\beta_1} X_1$

Dummy Variables for Linear Models

When creating a dummy variable with k categories, we create k-1 new variables.

- Marital Status has 3 categories, so we create 3-1=2 new variables. Here Single is the reference category.
  - Linear regression: Interpretation of coefficient estimate $\beta_{married}$: A married individual, on average, earns $\beta_{married}$ more than the reference category, a single individual.

*Note: The reference category is the category that does not appear as a dummy column after encoding.

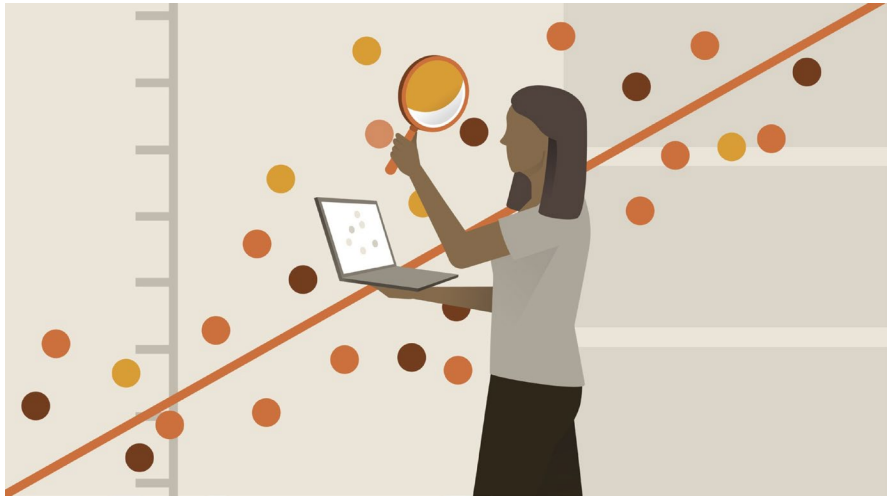
Income	Age	Marital Status
$45,000	23	Single
$48,000	25	Single
$54,000	24	Single
$57,000	29	Single
$65,000	38	Married
$69,000	36	Single
$78,000	40	Married
$83,000	59	Divorced
$98,000	56	Divorced
$104,000	64	Married
$107,000	53	Married

Income	Age	Married	Divorced
$45,000	23	0	0
$48,000	25	0	0
$54,000	24	0	0
$57,000	29	0	0
$65,000	38	1	0
$69,000	36	0	0
$78,000	40	1	0
$83,000	59	0	1
$98,000	56	0	1
$104,000	64	1	0
$107,000	53	1	0

Linear Regression Benefits and Assumptions

Linear Regression Benefits

- Simple model and readily interpretable



Main Linear Regression Assumption

- There is a linear relationship between response and predictor variables.

- Also, how can we determine which predictor variables are important with a large number of predictor variables?

Potential Problems with Linear Regression

There are several possible problems that one may encounter when fitting the linear regression model.

1. Non-linearity of the data
2. Correlation of the error terms
3. Non-constant variance of error terms
4. Outliers
5. High leverage points
6. Collinearity

See p. 92 of Introduction to Statistical Learning book.

*We will not focus on these since linear regression is not the main focus of this book. More complex models that we cover will alleviate most of these problems.

Potential Problems with Linear Regression: Multicollinearity

- What it is?
  - Two or more predictor variables are closely related to one another or highly correlated
  - Increases the variance of the coefficient estimates and makes the estimates very sensitive to minor changes in the model. The result is that the coefficient estimates are unstable and difficult to interpret.
- Ways to handle it
  1. The first is to drop one of the problematic variables from the regression.
     - This can usually be done without much compromise to the regression fit, since the presence of collinearity implies that the information that this variable provides about the response is redundant in the presence of the other variables.
     - If variable inflation factor (VIF) is greater than 5 or 10, then delete a variable with high VIF.
  2. The second solution is to combine the collinear variables together into a single predictor
     - Principal Components Analysis (unsupervised learning)

Predicting House Prices

Boston Housing Data variables

- **Medv** : median value of owner-occupied homes in $1000
- **crim** : per capita crime rate by town
- **zn** : proportion of residential land zoned for lots over 25,000 sq.ft.
- **Indus** : proportion of non-retail business acres per town
- **chas** : Charles River dummy variable (1 if tract bounds river; 0 otherwise)
- **nox** : nitrogen oxides concentration (parts per 10 million)
- **rm** : average number of rooms per dwelling
- **age** : proportion of owner-occupied units built prior to 1940
- **dis** : weighted mean of distances to five Boston employment centres
- **rad** : index of accessibility to radial highways
- **tax** : full-value property-tax rate per $10,000
- **ptratio**: pupil-teacher ratio by town
- **lstat** : lower status of the population (percent)

Can we predict median house price given data?