Intro to Machine Learning

Lecture 2

Adil Khan

a.khan@innopolis.ru

Recap

- What is machine learning?
- Why learn/estimate?
- Predictors and response variables
- Types of learning
- Regression and classification
- Parametric and non-parametric models
- Bias and variance

Today’s Objectives

- What is linear regression?
- Why study linear regression?
- What can we use it for?
- How to perform linear regression?
- How to estimate its performance?

We Will Start with this Example

TV	Radio	Newspaper	Sales
230.1	37.8	69.2	22.1
44.5	39.2	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	75.0	7.2

Advertising data:

Response (sales): in thousands of units sold

Predictors (TV, Radio, Newspaper): advertising budget in thousands of dollars

What we might want to know?

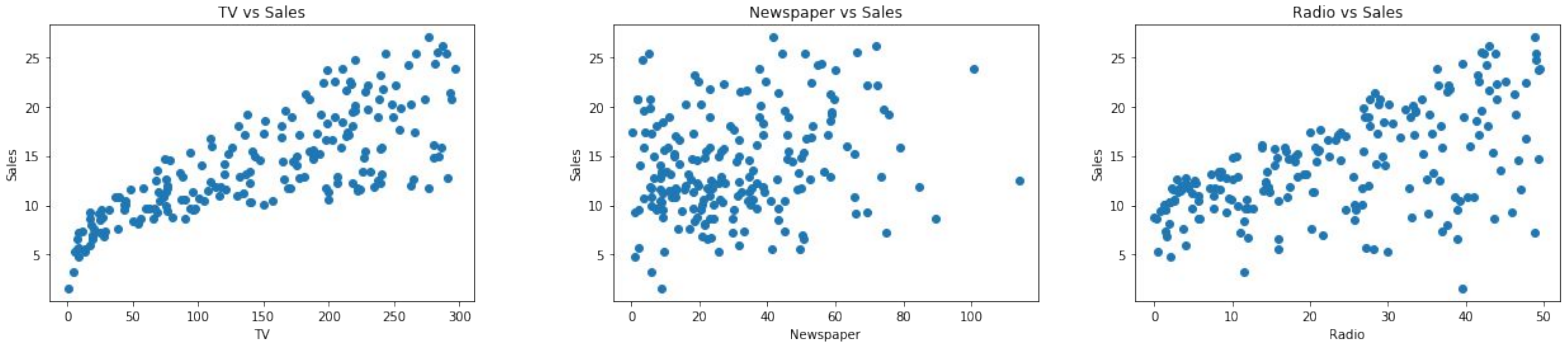
- Is there a relationship between advertising budget and sales?
- How strong is the relationship between advertising budget and sales?
- Which media contribute to sales?
- How accurately can we estimate the effect of each medium on sales?
- How accurately can we predict future sales?
- Is there synergy among the advertising media?

What we might want to know?

- 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 estimate the effect of each medium on sales?
- How accurately can we predict future sales?
- Is there synergy among the advertising media?

Prediction or
Inference?

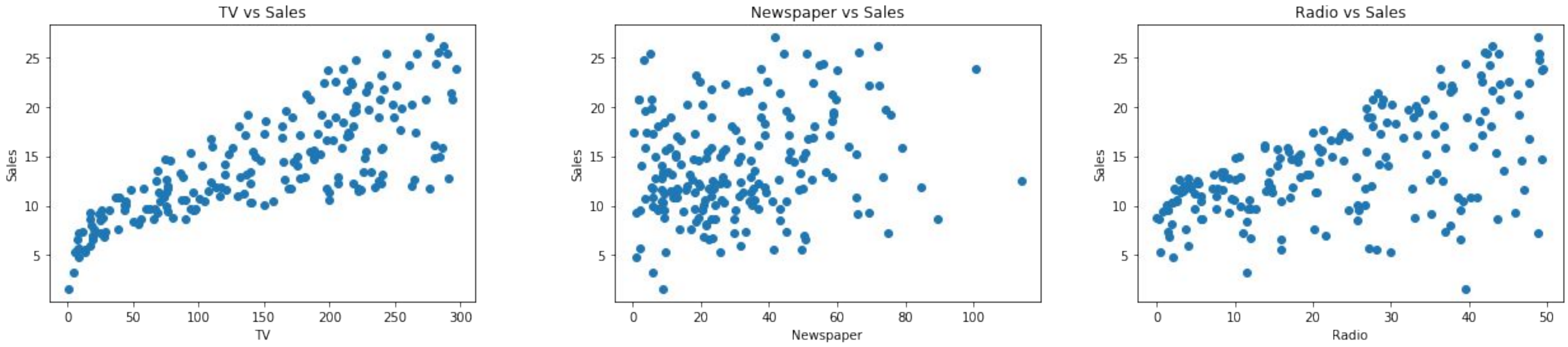
Formulate the Learning Problem



$$Sales = f(TV, Newspaper, Radio) + \epsilon$$

$$\widehat{Sales} \approx \hat{f}(TV, Newspaper, Radio)$$

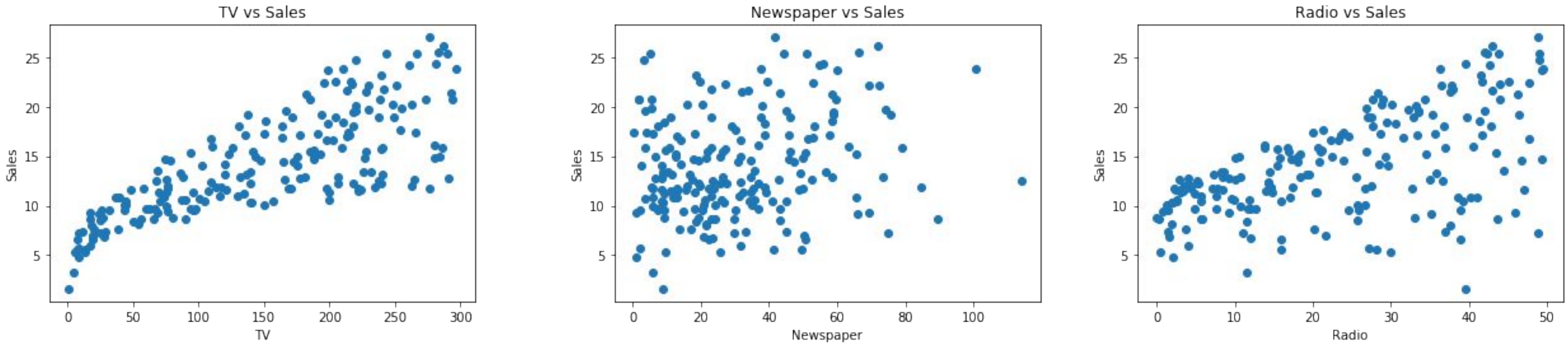
Determine the Nature of the Learning Problem



$$\widehat{Sales} \approx \hat{f}(TV, Newspaper, Radio)$$

Classification or Regression?

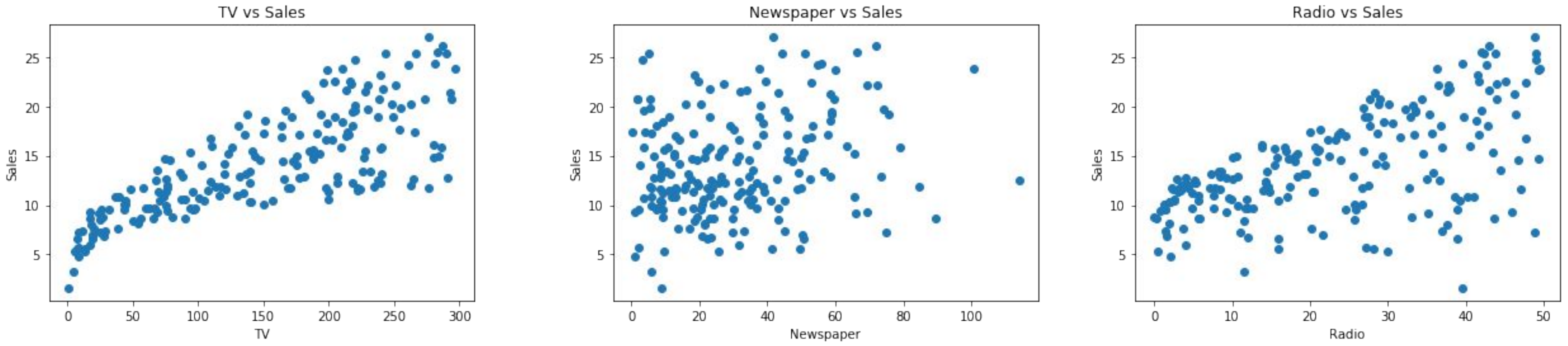
Simplify the Regression Problem



$$\widehat{Sales} \approx \hat{f}(TV, Newspaper, Radio)$$

Assume f to be a function of finite parameters

Further Simplify the Regression Problem



$$\widehat{Sales} \approx \hat{f}(TV, Newspaper, Radio)$$

Assume f to be a LINEAR function

Which Brings us to Linear Regression!

Linear Regression

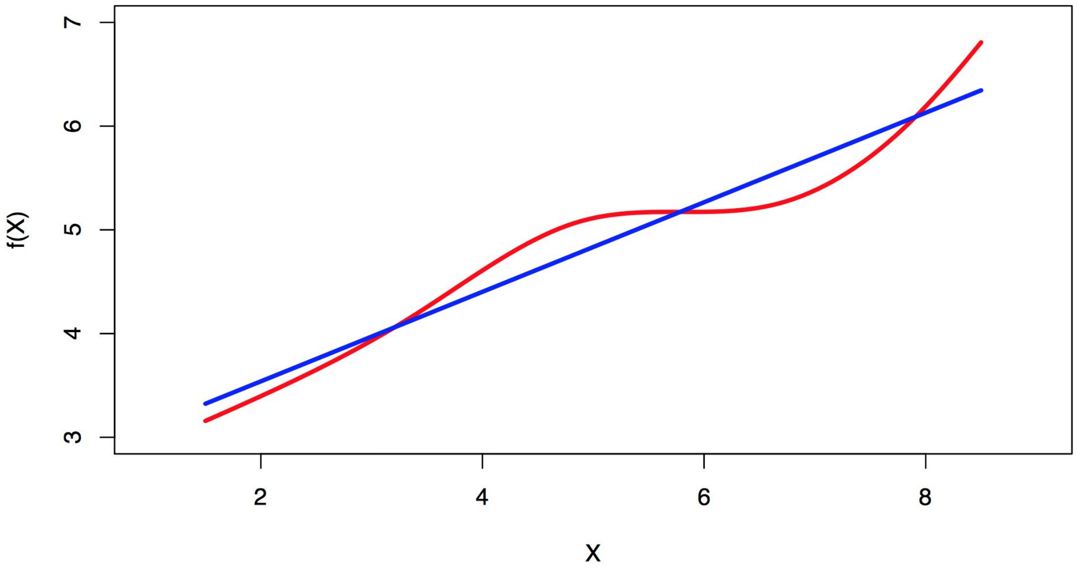
$$y = f(x) + \epsilon$$

$$\boxed{f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j .}$$

Linear Regression

- A simple supervised learning approach

- Assumes a linear relationship between the predictors and the response



$$Y = \beta_0 + \beta_1 X$$

Why study linear regression?

- Although it may seem overly simplistic, linear regression is extremely useful both conceptually and practically.

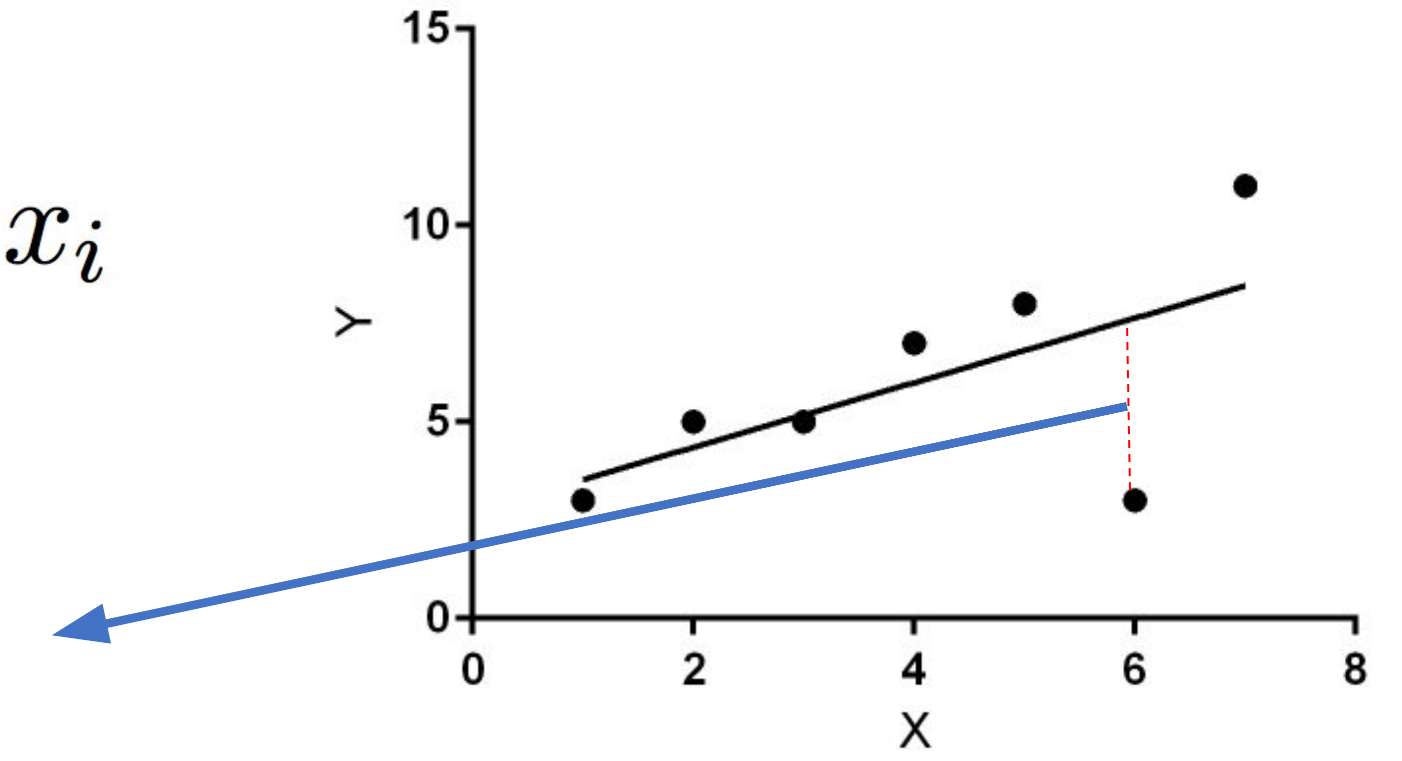
    - It is still a useful and widely used statistical learning method

    - It serves as a good jumping-off point for newer approaches:

Estimating LR Parameters by Least Squares (1)

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

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



Estimating Parameters by Least Squares (2)

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

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

- Residual sum of squares

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

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

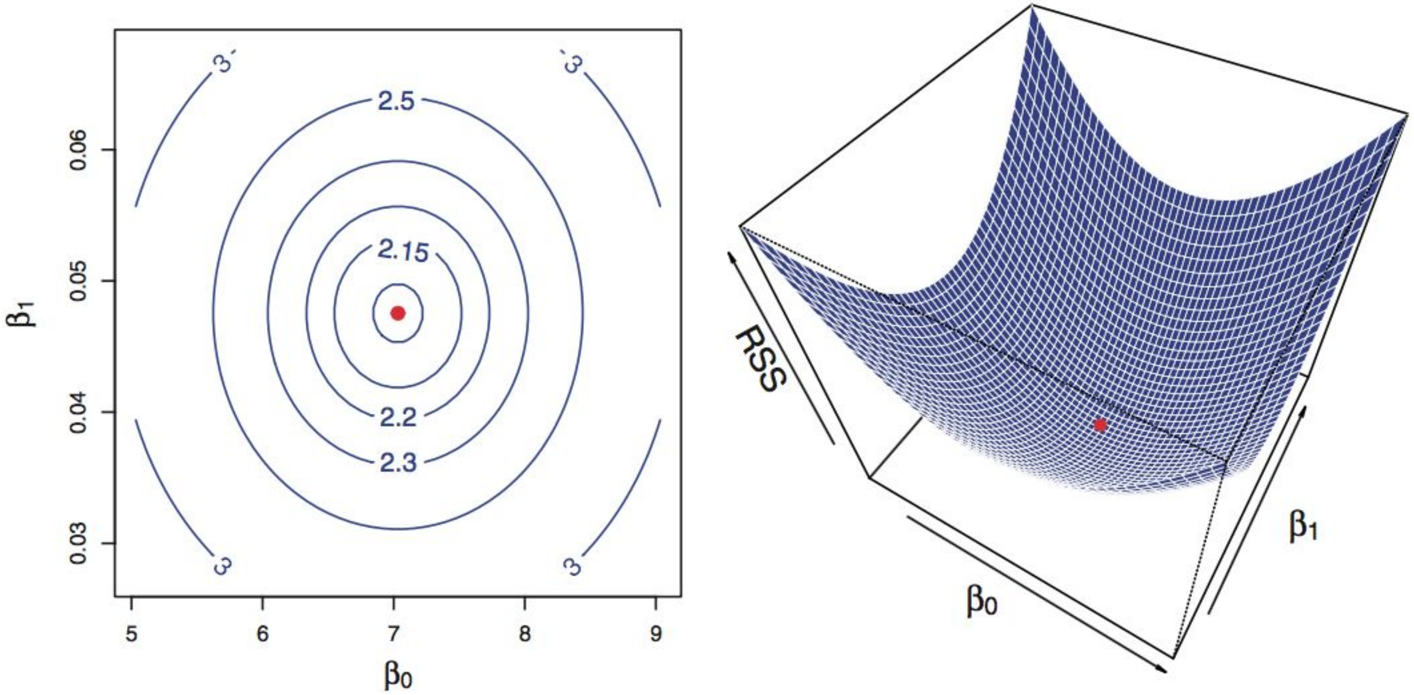
Estimating Parameters by Least Squares (3)

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

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

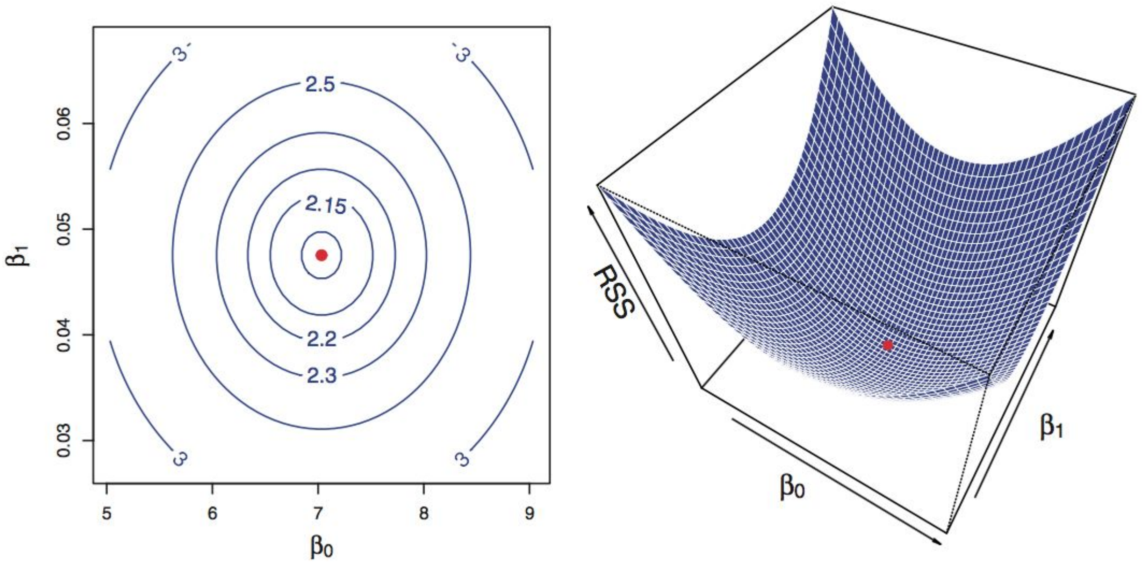
$$= \sum_{i=1}^{n} (y_i - \hat{\beta}_o - \hat{\beta}_1 x_i)^2$$

Estimating Parameters by Least Squares (4)



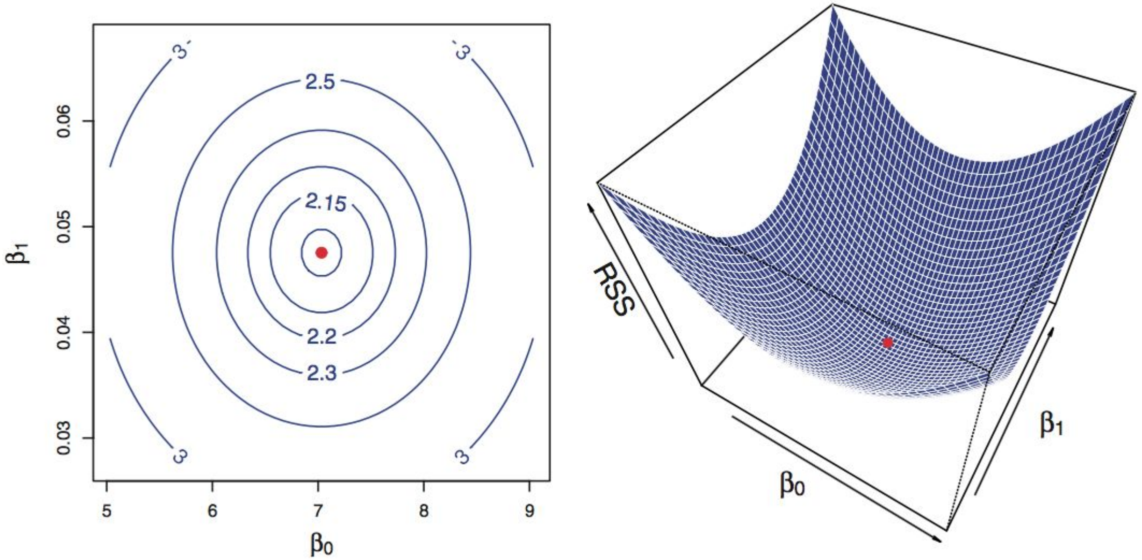
Contour and three-dimensional plots of the RSS

Estimating Parameters by Least Squares (5)



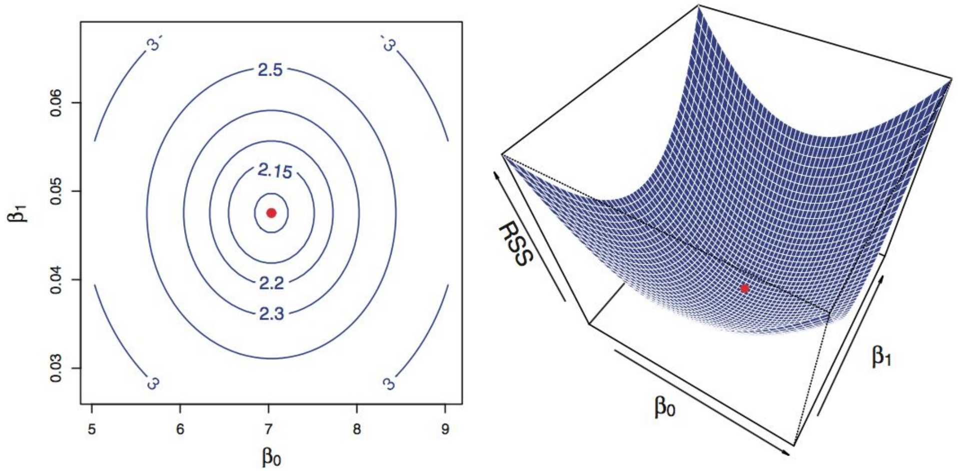
- Thus, we need to find values for our parameters that minimize the risk
- And, this is where the derivatives and gradients help us

Estimating Parameters by Least Squares (5)



- Thus,
  1. We will compute partial derivatives of RSS with respect to β_0 and β_1
  2. Set them to 0
  3. And solve for β_0 and β_1

Estimating Parameters by Least Squares (6)



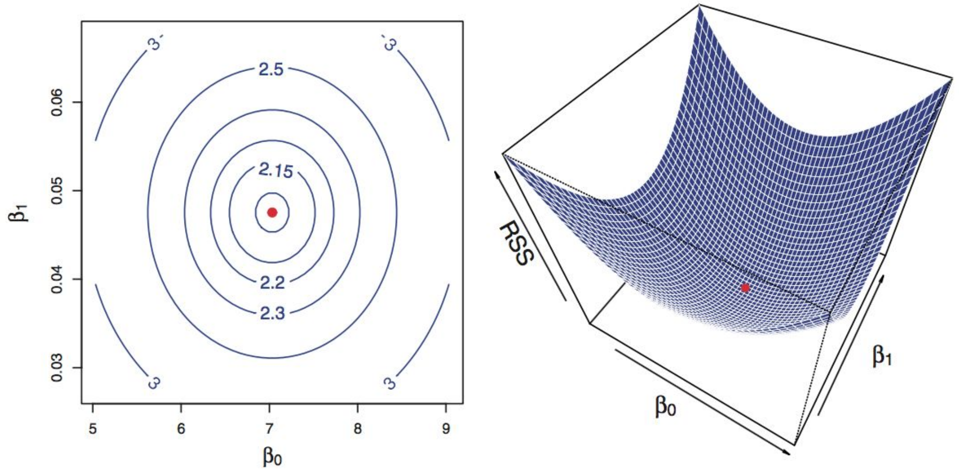
- Doing the said calculus and algebra, the minimizing values can be found as

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

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

where $\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i$ are the sample means.

See it for the Intercept. For ease I did not use the hat symbol



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

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

where $\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i$ are the sample means.

$$RSS = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

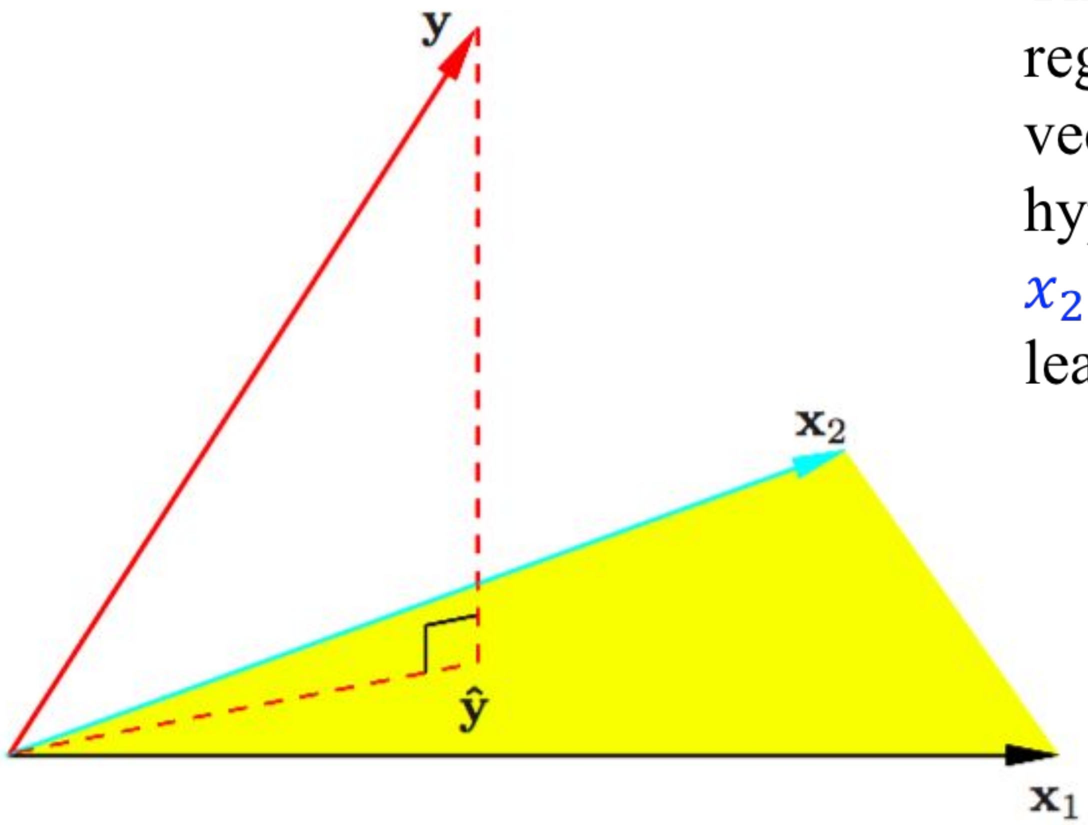
$$\frac{\partial \, RSS}{\partial \, \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)$$

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \beta_0 - \sum_{i=1}^{n} \beta_1 x_i = 0$$

$$\beta_0 = \frac{\sum_{i=1}^{n} y_i}{n} - \beta_1 \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

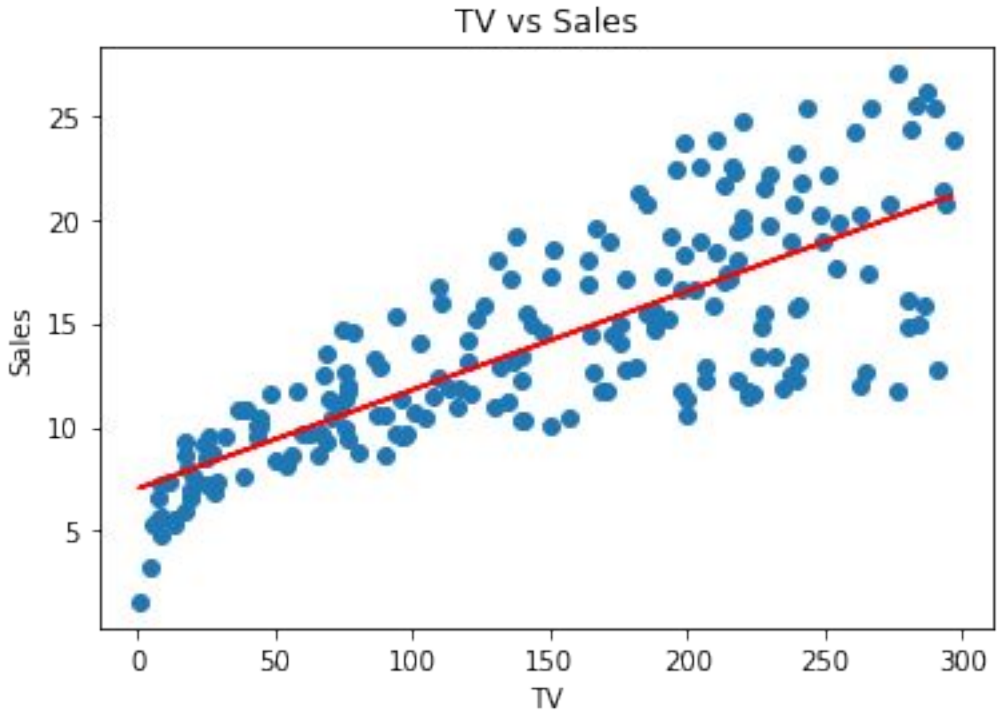
Geometry of Least Square Regression



The N-dimensional geometry of least squares regression with two predictors. The outcome vector y is orthogonally projected onto the hyperplane spanned by the input vectors $x_1$and x_2. The projection $\hat{y}$ represents the vector of the least squares predictions

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

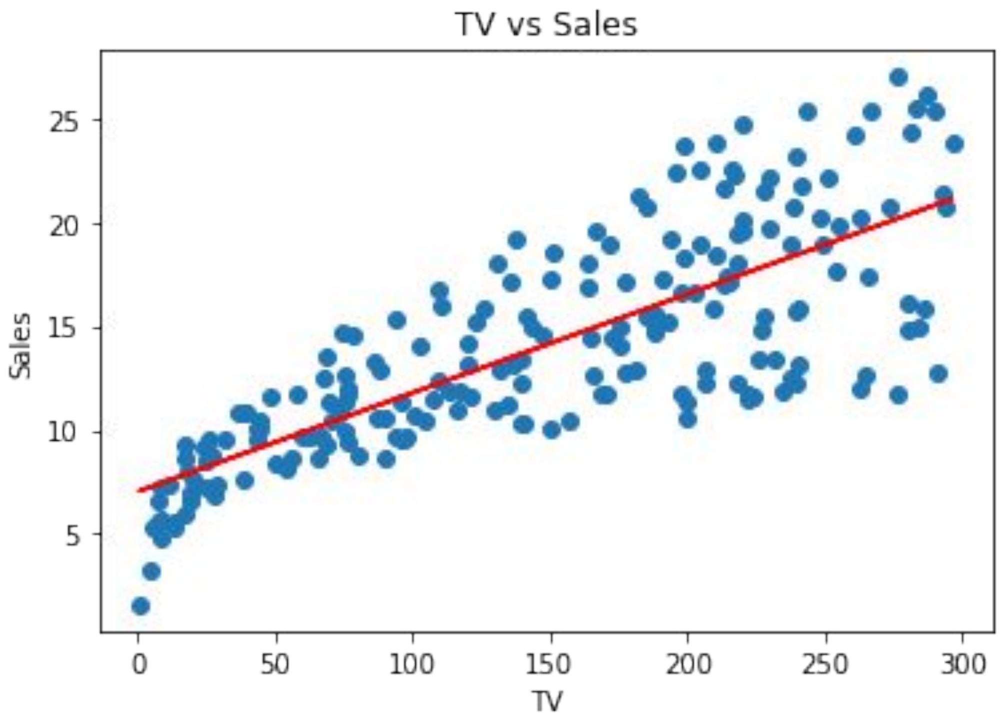
For our Sales Example



$$\widehat{Sales} \approx \hat{\beta}_0 + \hat{\beta}_1 TV$$

Parameters	Values
Intercept	7.0326
TV	0.475

Interpreting the Results



As per this estimation, an additional $1,000 spent on TV advertising is associated with selling approximately 47.5 additional units of the product.

Parameters	Values
Intercept	7.0326
TV	0.475

Now that we have the estimates, what is next?

- Goodness of fit
- Goodness of estimate

$$\widehat{Sales} \approx \hat{\beta}_0 + \hat{\beta}_1 TV$$

Parameters	Values
Intercept	7.0326
TV	0.475

Now that we have estimates, what is next?

- Goodness of fit (How best does the chosen model describe the data?)

- Goodness of estimate (Given the model, Is there really a relationship between response and predictor?)

$$\widehat{Sales} \approx \hat{\beta}_0 + \hat{\beta}_1 TV$$

Parameters	Values
Intercept	7.0326
TV	0.475

Goodness of Estimate (1)

- Is there really a relationship between sales (response) and TV (predictor)?

- Mathematically this corresponds to

$$H_0: \beta_1 = 0$$

  - verses

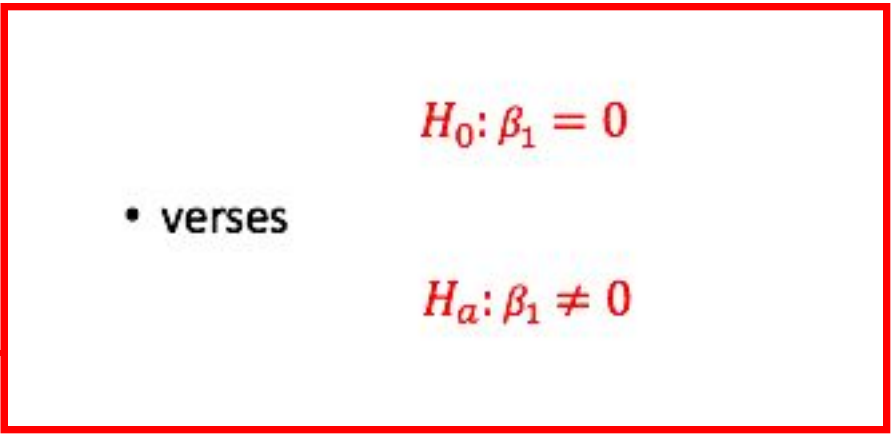
$$H_a: \beta_1 \neq 0$$

$$\widehat{Sales} \approx \hat{\beta}_0 + \hat{\beta}_1 TV$$

Parameters	Values
Intercept	7.0326
TV	0.475

Goodness of Estimate (2)

- Is there really a relationship between sales (response) and TV (predictor)?



$H_0: \beta_1 = 0$

- verses

$H_a: \beta_1 \neq 0$

- For this, we calculate t-statistics

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

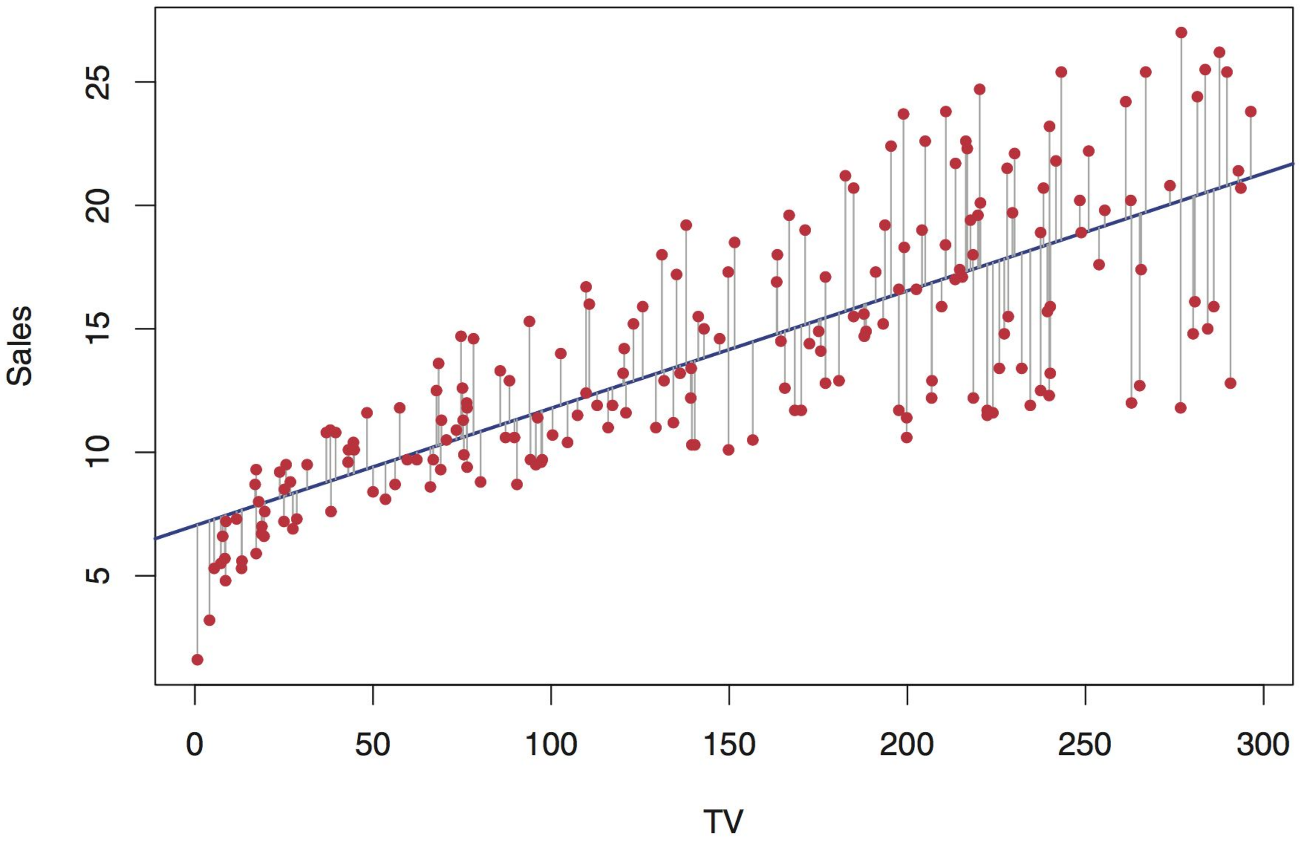
$$\widehat{Sales} \approx \hat{\beta}_0 + \hat{\beta}_1 TV$$

- Where SE is an estimate of how close the estimated parameter value is to its true value

Parameters	Values
Intercept	7.0326
TV	0.475

Aside: SE

$$\mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$



For Our Example

- t-statistics

The greater the magnitude of t, the greater the evidence against the null hypothesis

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

$$\widehat{Sales} \approx \hat{\beta}_0 + \hat{\beta}_1 TV$$

Parameters	Values	t-value
Intercept	7.0326	15.360
TV	0.475	17.668

For Our Example

- t-statistics

The greater the magnitude of t, the greater the evidence against the null hypothesis

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

Remember, we are dealing with estimates, thus we should also eliminate the risk that the resulting t-value was not by chance.

$$\widehat{Sales} \approx \hat{\beta}_0 + \hat{\beta}_1 TV$$

Parameters	Values	t-value
Intercept	7.0326	15.360
TV	0.475	17.668

Chances of getting the Resulting t-value

- For this, we calculate *p-value*
  - Probability of getting $|t|$ assuming β_1 was 0

$$\widehat{Sales} \approx \hat{\beta}_0 + \hat{\beta}_1 TV$$

Parameters	Values	t-value	p-value
Intercept	7.0326	15.360	< 0.0001
TV	0.475	17.668	< 0.0001

Was our Assumption about the Model Correct?

- What is the extent to which the model fits the data?

- This can judged using R^2 *statistics*

$$\widehat{Sales} \approx \hat{\beta}_0 + \hat{\beta}_1 TV$$

Parameters	Values	t-value	p-value
Intercept	7.0326	15.360	< 0.0001
TV	0.475	17.668	< 0.0001

R^2

R-squared: how much do we gain by using the *learned models* instead of using the mean as the model (no independent variables)

$$\text{TSS} = \sum (y_i - \bar{y})^2$$

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

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

For Our Example

- R^2 *statistics*

- In this case, it is 0.612

$$\widehat{Sales} \approx \hat{\beta}_0 + \hat{\beta}_1 TV$$

Parameters	Values	t-value	p-value
Intercept	7.0326	15.360	< 0.0001
TV	0.475	17.668	< 0.0001

Multiple Linear Regression (1)

- Simple linear regression is a useful approach for predicting a response on the basis of a single predictor variable.

- However, in practice we often have more than one predictor

    - Sales (TV, Radio, Newspaper)
    - Income (Years of education, Years of experience, Age, Gender)

Multiple Linear Regression (2)

- Options

1. Fit p separate linear regressions (where p is the number of predictors)

2. Extend the simple linear regression model, so that it can directly accommodate multiple predictors

Multiple Linear Regression (3)

- Options

1. Fit p separate linear regressions (where p is the number of predictors)

2. Extend the simple linear regression model, so that it can directly accommodate multiple predictors

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

Multiple Linear Regression (4)

- For p predictors,

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

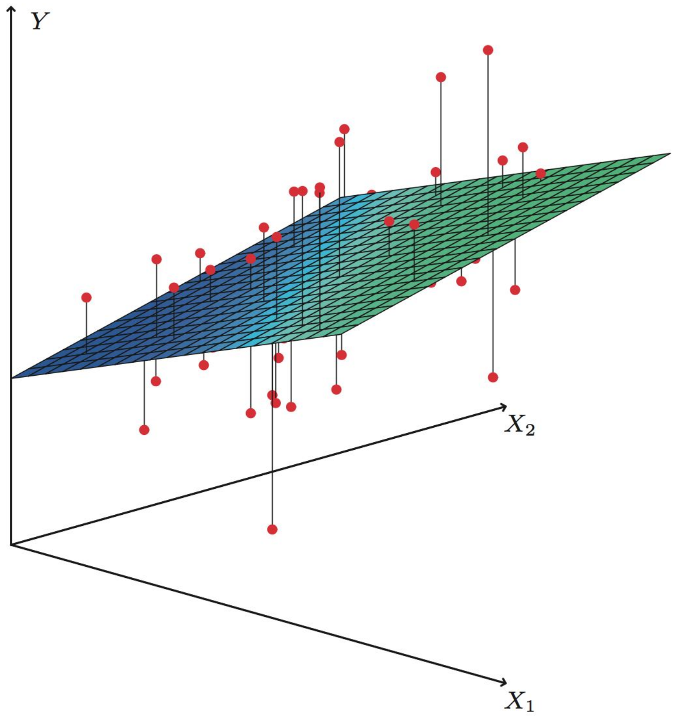
- The parameters are estimated using the same least squares approach that we saw in the context of simple linear regression

$$\mathrm{RSS}(\beta) = \sum_{i=1}^{N} \Big(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \Big)^2.$$

- The coefficients can be calculated using statistical packages

Multiple Linear Regression (5)

- For two predictors, the regression might look as follows



For Our Sales Example

Parameters	Values	t-value	p-value
Intercept	2.939	9.42	< 0.0001
TV	0.46	32.81	< 0.0001
Radio	0.189	21.89	< 0.0001
Newspaper	-0.001	-0.18	< 0.8599

$$\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{newspaper} + \epsilon.$$

For the Advertising data, least squares coefficient estimates of the multiple linear regression of number of units sold on radio, TV, and newspaper advertising budgets.

Multiple Linear Regression (7)

Parameters	Values	t-value	p-value
Intercept	2.939	9.42	< 0.0001
TV	0.46	32.81	< 0.0001
Radio	0.189	21.89	< 0.0001
Newspaper	-0.001	-0.18	< 0.8599

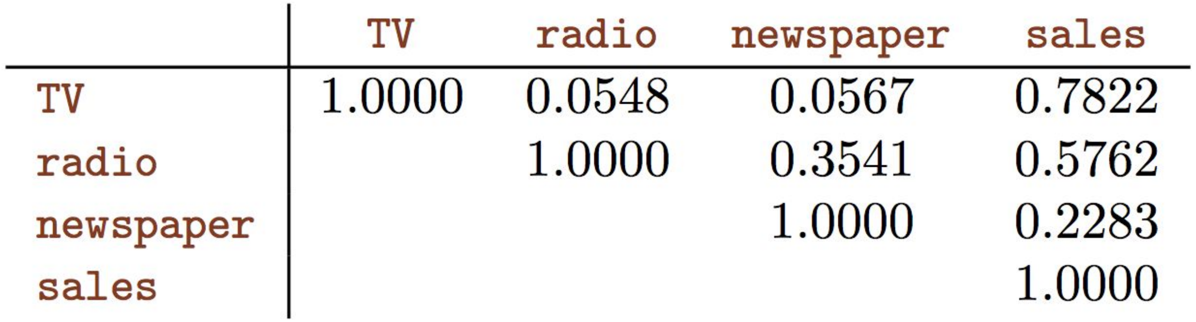
Compare the results for ‘Newspaper’ of **multiple regression (above)** to that of **linear regression** (above)

Parameters	Values	t-value	p-value
Intercept	12.351	19.88	< 0.0001
Newspaper	0.055	3.30	0.00115

Multiple Linear Regression (7)

Parameters	Values	t-value	p-value
Intercept	2.939	9.42	< 0.0001
TV	0.46	32.81	< 0.0001
Radio	0.189	21.89	< 0.0001
Newspaper	-0.001	-0.18	< 0.8599

Correlation matrix for TV, radio, newspaper, and sales for the Advertising data



	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

Interpreting the Results of MLR (1)

- 1. Is there any predictor which is useful in predicting the response?
  - We might think that (just like LR) we can use p-value for this, but **we are wrong**

Parameters	Values	t-value	p-value
Intercept	2.939	9.42	< 0.0001
TV	0.46	32.81	< 0.0001
Radio	0.189	21.89	< 0.0001
Newspaper	-0.001	-0.18	< 0.8599

Interpreting the Results of MLR (2)

- 1. Is there any predictor which is useful in predicting the response?
  - Thus we use another measure called F-statistics

$$F = \frac{(\mathrm{TSS} - \mathrm{RSS})/p}{\mathrm{RSS}/(n - p - 1)}$$

These two quantities are expected to be the same under ***Null Hypothesis***

Interpreting the Results of MLR (3)

- 1. Is there any predictor which is useful in predicting the response?

  - Thus we use another measure called F-statistics

Parameters	Values	t-value	p-value
Intercept	2.939	9.42	< 0.0001
TV	0.46	32.81	< 0.0001
Radio	0.189	21.89	< 0.0001
Newspaper	-0.001	-0.18	< 0.8599

F-statistics	570

Since this is far larger than 1, it provides compelling evidence against the null hypothesis H0.
In other words, the large F-statistic suggests that at least one of the advertising media must be related to sales

Interpreting the Results of MLR (4)

- 1. Is there any predictor which is useful in predicting the response?

  - But how far away from 0 F-statistics has to be?

Interpreting the Results of MLR (5)

- 2. Do all the predictors help explain the response or is only a subset of them useful?

  □ Forward selection

  □ Backward selection

  □ Mixed selection

Do all the predictors help explain the response or is only a subset of them useful?

- 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.

Do all the predictors help explain the response or is only a subset of them useful?

- Backward Selection

  - We start with all variables in the model, and remove the variable with the largest p-value—that is, the variable that is the least statistically significant.

  - 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. For instance, we may stop when all remaining variables have a p-value below some threshold.

Do all the predictors help explain the response or is only a subset of them useful?

- Mixed Selection

□ Left as home reading

Interpreting the Results of MLR (6)

- 3. How well does the model fit the data?

  □ Same as LR with single parameter (R-squared)

Potential Problems with Linear Regression

- Non-linearity of f
- Correlation of error terms
- Non-constant variance of error terms
- Outliers
- High-leverage points
- Collinearity

Did we achieve today’s objectives objectives?

- What is linear regression?
- Why study linear regression?
- What can we use it for?
- How to perform linear regression?
- How to estimate its performance?