Statistical Learning for Data Science

Lecture 08

唐晓颖

电子与电气工程系

南方科技大学

March 15, 2023

Simple Linear Regression

❑ Assessing the accuracy of the coefficient estimates

Hypothesis testing

To test the null hypothesis, we need to determine whether $\hat{\beta}_1$, our estimate for β_1 , is sufficiently far from zero that we can be confident that $\hat{\beta}_1$ is non-zero.

How far is far enough?

It depends on the accuracy of $\hat{\beta}_1$ (depends on $SE(\hat{\beta}_1)$). If $SE(\hat{\beta}_1)$ is small, then even relatively small values of $\hat{\beta}_1$ may provide strong evidence that $\hat{\beta}_1 \neq 0$, and hence there is a relationship between X and Y. In contrast, if $SE(\hat{\beta}_1)$ is large, then $\hat{\beta}_1$ must be large in absolute value in order for us to reject the null hypothesis.

Simple Linear Regression

- ❑ Assessing the accuracy of the coefficient estimates

Hypothesis testing

How far is far enough?

t-statistic (to measure the number of standard deviations that $\hat{\beta}_1$ is away from 0)

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

Note: in every Hypothesis Testing case, we need to define a testing statistic!

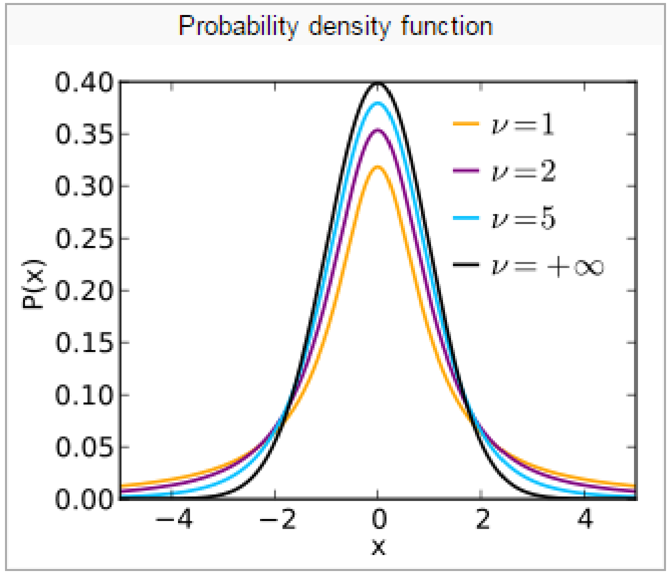
Simple Linear Regression

- ❑ Assessing the accuracy of the coefficient estimates

Hypothesis testing

When there is no relationship between *X* and *Y*, namely $\beta_1 = 0$, we have

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)} \sim \text{t-distribution with df } = n - 2$$



p-value: $\Pr(T \geq |t|)$

A small p-value indicates that it is ***unlikely*** to observe such a substantial association between the predictor and the response ***due to chance***, in the absence of any real association between the predictor and the response.

Simple Linear Regression

❑ Assessing the accuracy of the coefficient estimates

Hypothesis testing

We *reject the null hypothesis* – that is, we declare a relationship to exist between *X* and *Y* – if the p-value is small enough (typically used p-value cutoffs are 0.05 or 0.01).

Results for the advertising data (sales-versus-TV):

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

Simple Linear Regression

❑ Assessing the accuracy of the coefficient estimates

Hypothesis testing

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

$\beta_1 \neq 0 \Rightarrow$ there is a (statistically significant) relationship between TV and sales

$\beta_0 \neq 0 \Rightarrow$ In the absence of TV expenditure, sales are (statistically significantly) non-zeros

Simple Linear Regression

❑ Assessing the accuracy of the coefficient estimates

To quantify the extent to which the model fits the data

Residual Standard Error (RSE)

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

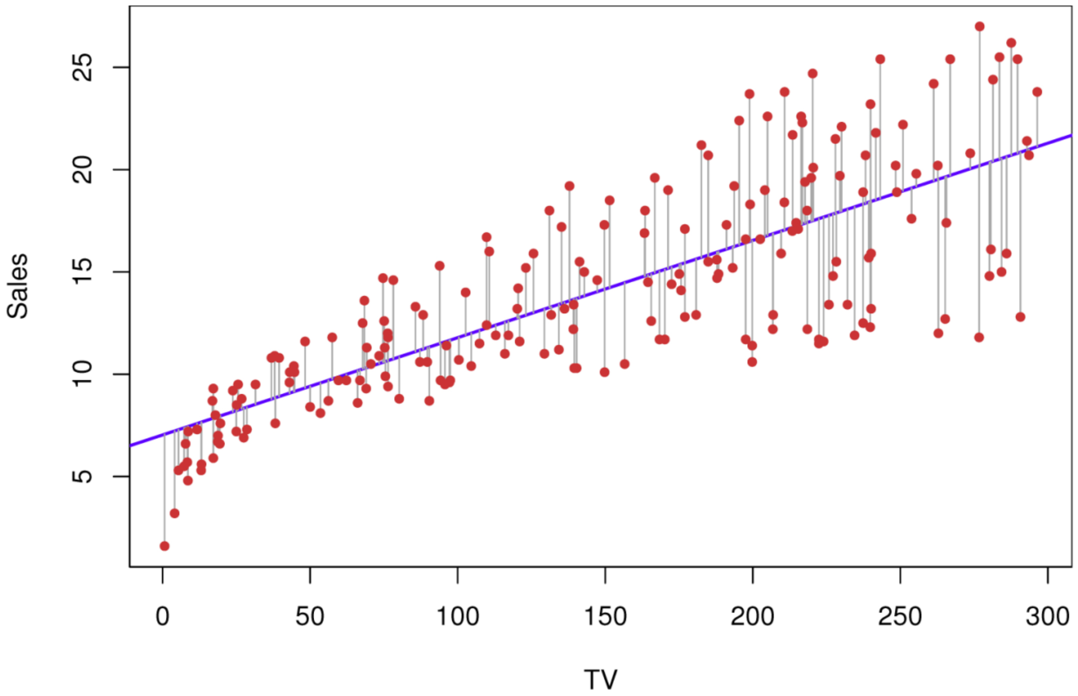
The RSE is an estimate of the standard deviation of ε . Roughly speaking, it is the average amount that the response will deviate from the true regression line.

Simple Linear Regression

❑ Assessing the accuracy of the coefficient estimates

To quantify the extent to which the model fits the data

Residual Standard Error (RSE)



$\mathrm{RSE}: 3.26$

A measure of the *lack of fit* of the model

How to interpret?

- Actual sales in each market deviate from the true regression line by approximately 3,260 units, on average.

- Even if the model were correct and the true values of the unknown coefficients were known exactly, any prediction of sales on the basis of TV advertising would still be off by about 3,260 units on average.

Simple Linear Regression

❑ Assessing the accuracy of the coefficient estimates

To quantify the extent to which the model fits the data

R^2 *Statistic*

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}}, \text{ with TSS} = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

- ○ TSS measures the total variation in the response *Y*, is the amount of variability inherent in the response before the regression is performed.
- ○ RSS measures the amount of variability that is left unexplained after performing the regression.
- ○ TSS-RSS measures the amount of variability in the response that is explained by performing the regression.
- ○ R^2 measures the proportion of variability in Y that can be explained using X.

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

Simple Linear Regression

❑ Assessing the accuracy of the coefficient estimates

To quantify the extent to which the model fits the data

R^2 *Statistic*

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}}, \text{ with TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- Ranges between 0 and 1.
- Independent of the scale of Y.
- A number that is close to 1 indicates that a large proportion of the variability in the response has been explained by the regression.
- A number near 0 indicates that the regression did not explain much of the variability in the response (the linear model is wrong, the inherent error is high).

Simple Linear Regression

❑ Assessing the accuracy of the coefficient estimates

To quantify the extent to which the model fits the data

R^2 *Statistic*

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}}, \text{ with } \mathrm{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

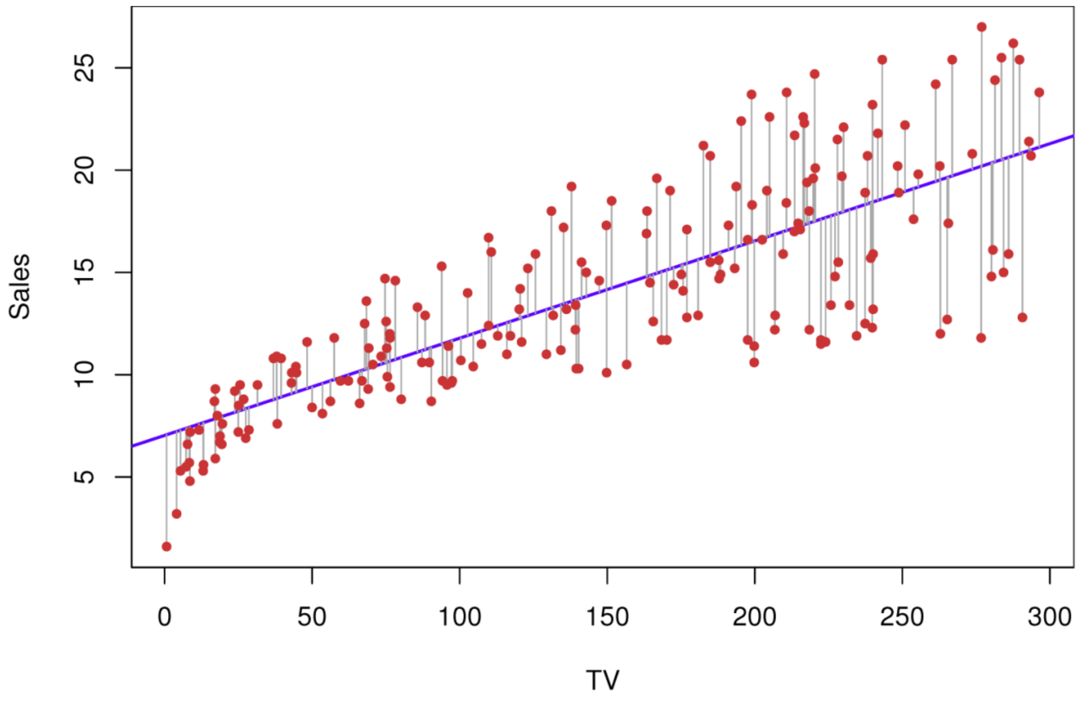
- It can only be used to measure the fitting degree of the linear regression model, not the nonlinear regression model
- It cannot be used to compare the performance of different sets of independent variables in building regression models because different sets of independent variables can have an impact on the R^2 value.

Simple Linear Regression

- ❑ Assessing the accuracy of the coefficient estimates

To quantify the extent to which the model fits the data

R^2 *Statistic*



$R^2 : 0.61$

How to interpret?

- Just under two-thirds of the variability in sales is explained by a linear regression on TV.

Simple Linear Regression

❑ Assessing the accuracy of the coefficient estimates

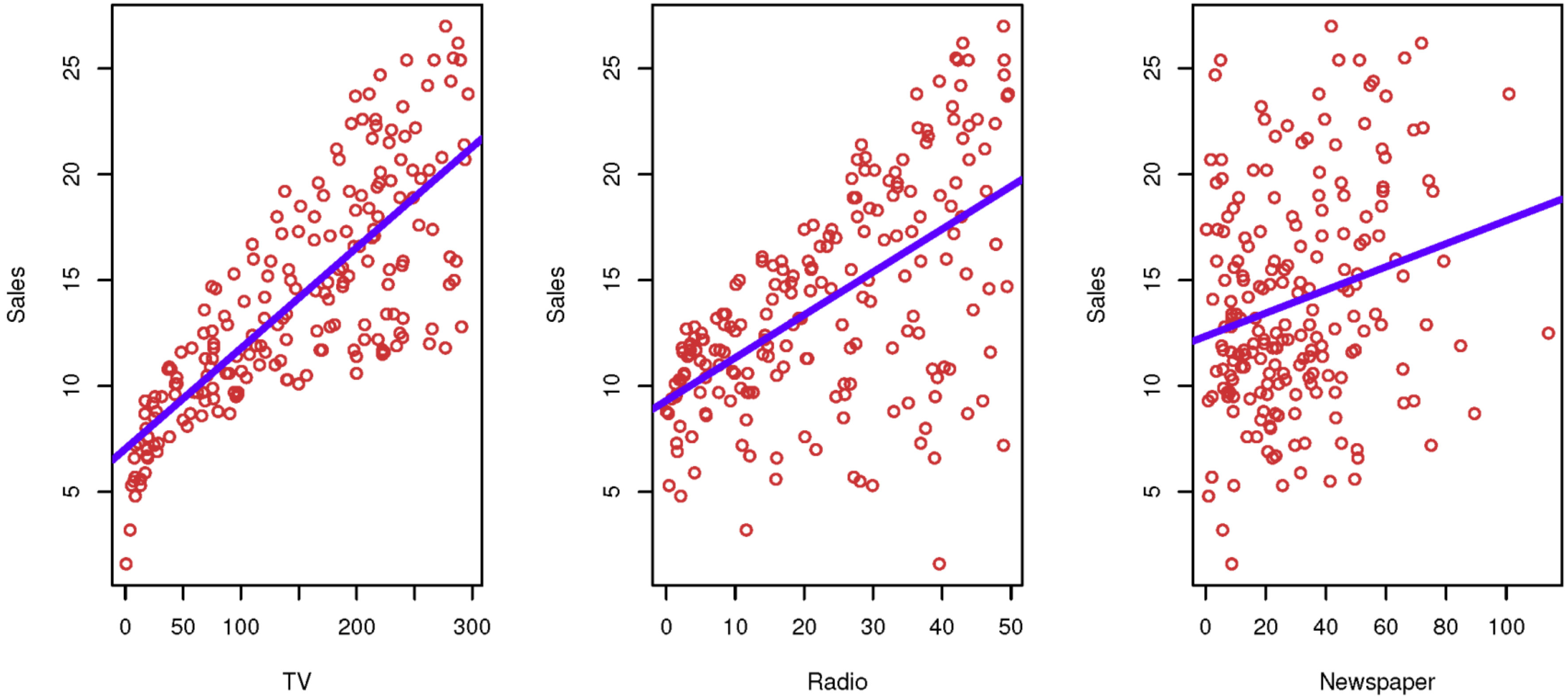
To quantify the extent to which the model fits the data

$$R^2 = [\mathrm{cor}(X, Y)]^2$$

The above equation is only true for simple linear regression, but not for the multiple linear regression case.

Multiple Linear Regression

- ❑ In practice, we often have more than one predictor.



Multiple Linear Regression

- ❑ Suppose we have p distinct predictors

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

- ○ X_j represents the *j*th predictor.
- ○ β_j quantifies the association between X_j and the response (the average effect on the response of a one unit increase in the predictor, *holding all other predictors fixed*).

- ❑ Advertising data

$$\text{sales} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \beta_3 \times \text{newspaper} + \varepsilon$$

Multiple Linear Regression

- Interpreting regression coefficients
  - The ideal scenario is when the predictors are uncorrelated (a balanced design):
    - Each coefficient can be estimated and tested separately.
    - Interpretations such as, "a unit change in X_j is associated with a β_j change in Y, while *all the variables stay fixed*", are possible.
  - Correlations among predictors cause problems:
    - The variance of all coefficients tends to increase, sometimes dramatically
    - Interpretations become hazardous --- when X_j changes, everything else changes.
  - *Claims of causality* should be avoided for observational data.

Multiple Linear Regression

- The woes of (interpreting) regression coefficients
  - A regression coefficient β_j estimates the expected change in Y per unit change in X_j *with all other predictors held fixed*. But predictors usually change together!
  - Example: *Y* = number of tackles by a football player in a season; *W* and *H* are his weight and height. Fitted regression model is $\hat{Y} = b_0 + 0.5W - 0.1H$. How do we interpret $\hat{\beta}_2 = -0.1 < 0$?

Multiple Linear Regression

- Two quotes by famous statisticians

“Essentially, all models are wrong, but some are useful”

– by George Box

“The only way to find out what will happen when a complex system is disturbed is to disturb the system, not merely to observe it passively”

– by Fred Mosteller and John Tukey, paraphrasing George Box

Multiple Linear Regression

- Estimating the Regression Coefficients

  - Given estimates $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p$, we can make predictions using the formula

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

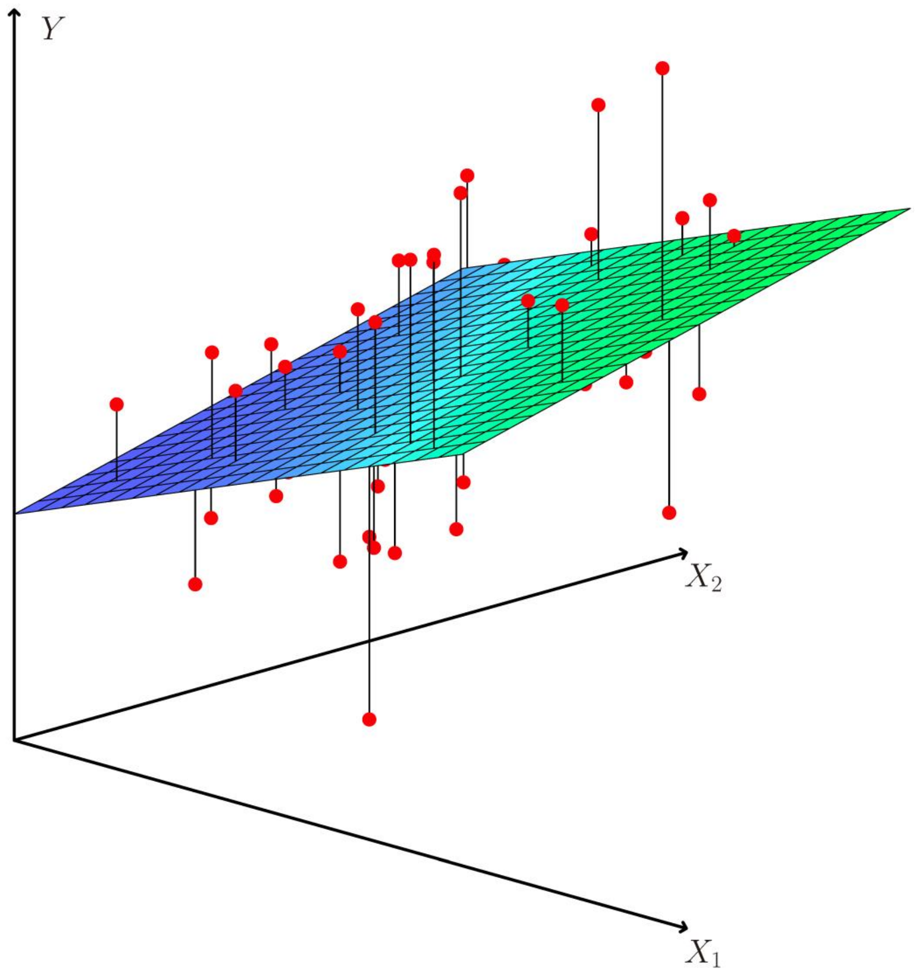
  - Similar to simple linear regression, we estimate $\beta_0, \beta_1, ..., \beta_p$ as the values that minimize the RSS

$$\begin{aligned} \mathrm{RSS} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_p x_{ip})^2 \end{aligned}$$

  - The values $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p$ that minimize the above RSS are the multiple least squares regression coefficient estimates.

Multiple Linear Regression

- Estimating the Regression Coefficients



An example with two predictors and one response. In this case, the least squares regression line becomes a plane. The plane is chosen to minimize the sum of the squared vertical distances between each observation and the plane.

Linear Regression

- Simple Linear Regression & Multiple Linear Regression

	Simple Linear Regression	Multiple Linear Regression
Independent Variables	Only one	Multiple
Model form	$Y = \beta0 + \beta1X$	$Y = \beta0 + \beta1X1 + \beta2X2 + ... + \beta pXp$
Objective	Describing the relationship between two variables and predicting the dependent variable	Describing the relationship between multiple variables and predicting the dependent variable
Evaluation metrics	Residual sum of squares (SSE), coefficient of determination (R^2), T statistic, standard error (SE), etc.	Residual sum of squares (SSE), coefficient of determination (R^2), F statistic, standard error (SE), etc.
Model fitting	Least squares method	Least squares method

Linear Regression

- Simple Linear Regression & Multiple Linear Regression

  - *Multiple Linear Regression: can be considered as an extension of simple linear regression*
  - *Both are used to describe the relationship between independent variables and dependent variables, but multiple linear regression can handle the influence of multiple independent variables on the dependent variable*
  - *Multiple linear regression can be applied to more practical problems*

Multiple Linear Regression

- Estimating the Regression Coefficients

Advertising data

	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

Interpretation: For a given amount of TV and newspaper advertising, spending an additional $1,000 on radio advertising leads to an increase in sales by approximately 189 units.

Multiple Linear Regression vs. Multiple simple linear regressions

	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

	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

	Coefficient	Std. Error	t-statistic	p-value
Intercept	9.317	0.562	16.54	< 0.0001
radio	0.203	0.020	9.92	< 0.0001

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

The simple and multiple regression coefficients can be quite different!

Multiple Linear Regression vs. Multiple simple linear regressions

The multiple regression suggests no relationship between sales and newspaper (p=0.8599) while the simple linear regression suggests the opposite (p<0.0001).

Why?

Correlation matrix

	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

Suppose the multiple regression is correct and newspaper advertising has no direct impact on sales, but radio advertising does increase sales. 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 regression examining only sales versus newspaper, we will observe that **higher values of newspaper tend to be associated with higher values of sales**.

Multiple Linear Regression vs. Multiple simple linear regressions

The multiple regression suggests no relationship between sales and newspaper while the simple linear regression suggests the opposite.

Why?

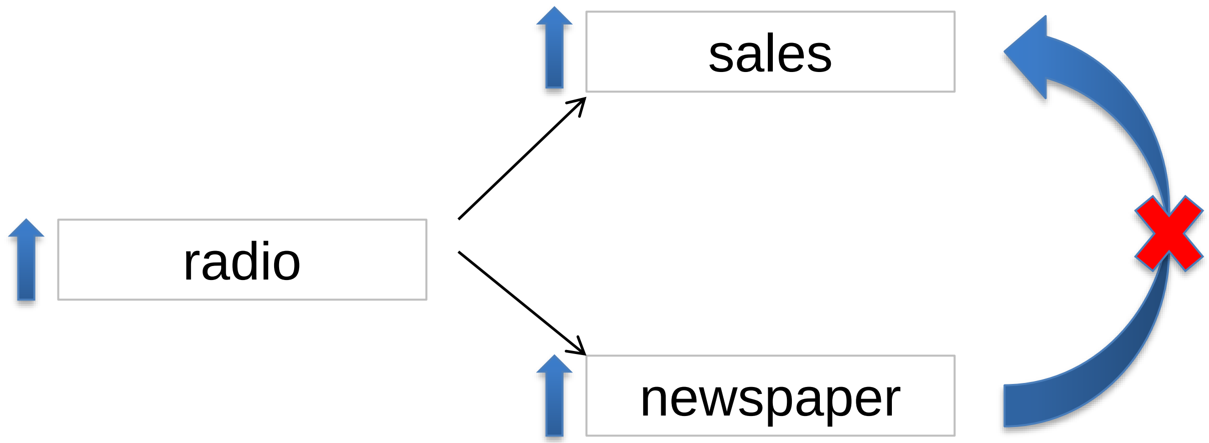
Correlation matrix

	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

Reason: newspaper gets “credit” for the effect of radio on sales.

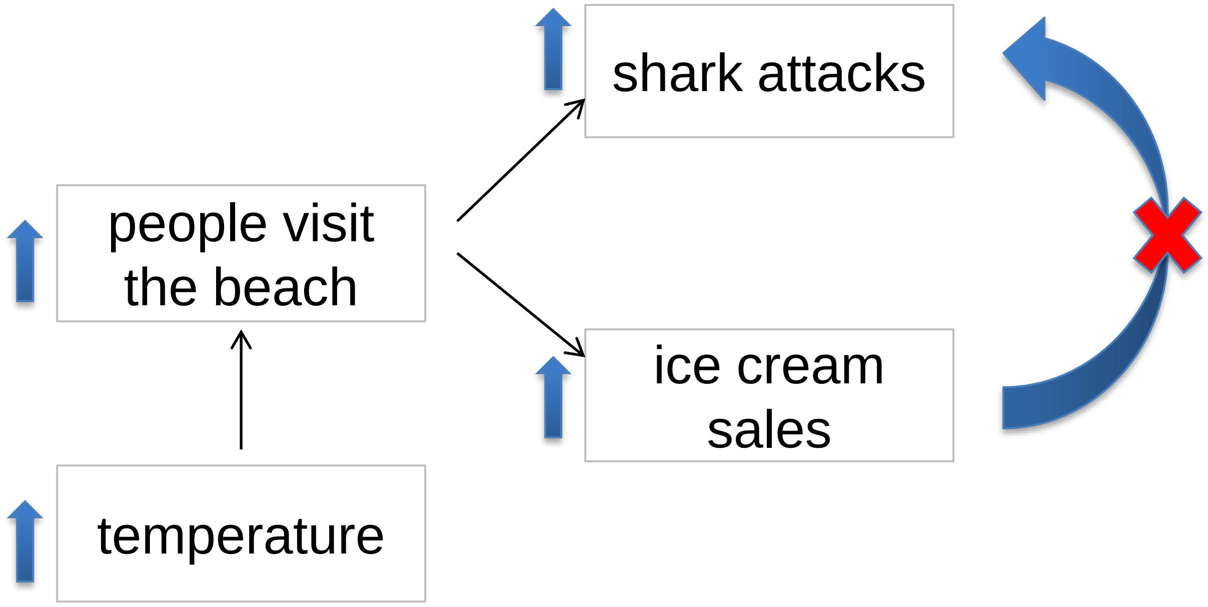
Multiple Linear Regression vs. Multiple simple linear regressions

The multiple regression suggests no relationship between sales and newspaper while the simple linear regression suggests the opposite.



Multiple Linear Regression vs. Multiple simple linear regressions

This slightly counterintuitive result is very common in many real life situations.



However, running a regression of shark attacks versus ice cream sales do show a positive correlation, similar to that seen between sales and newspaper.

Multiple Linear Regression

- Some important questions

1. Is there a relationship between the response and predictors?

2. Do all the predictors help to explain Y, or is only a subset of the predictors useful?

3. How well does the model fit the data?

4. Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

Multiple Linear Regression

1. Is there a relationship between the response and predictors?

Hypothesis testing $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$

Testing the *null hypothesis* of

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$$

Versus the *alternative hypothesis*

$$H_a : \text{ at least one } \beta_j \text{ is non-zero}$$

Multiple Linear Regression

1. Is there a relationship between the response and predictors?

Hypothesis testing $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$

We use the *F-statistic*

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

with

$$\mathrm{TSS} = \sum_{i=1}^{n} (y_i - \overline{y})^2, \mathrm{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Multiple Linear Regression

1. Is there a relationship between the response and predictors?

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

We have $E\{RSS/(n-p-1)\} = \sigma^2$

PROOF

When the null hypothesis is true, we have $E\{(TSS - RSS)/p\} = \sigma^2$

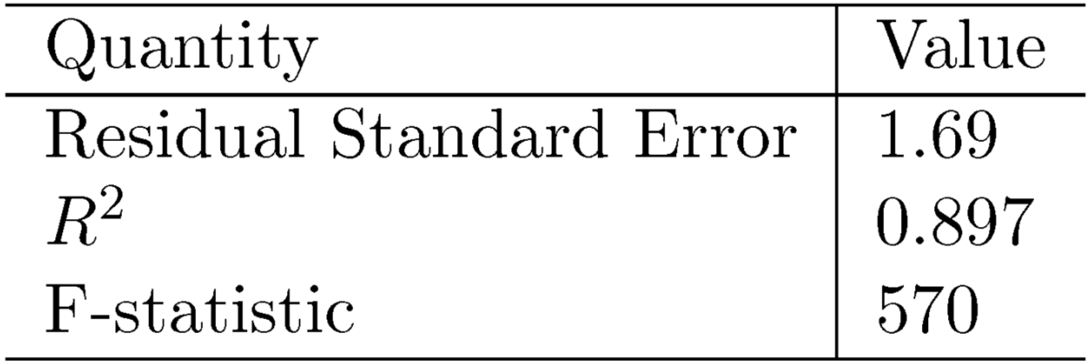
⬇

When there is no relationship between the response and predictors, one would expect the F-statistic to take on a value close to 1.

Multiple Linear Regression

1. Is there a relationship between the response and predictors?

Advertising data



Quantity	Value
Residual Standard Error	1.69
R^2	0.897
F-statistic	570

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

Multiple Linear Regression

1. Is there a relationship between the response and predictors?

How large does the F-statistic need to be before we can reject the null hypothesis and conclude that there is a relationship?

It depends on the values of n and p

- ❖ When n is large, an F-statistic that is just a little larger than 1 might still provide evidence against the null hypothesis
- ❖ When n is small, a larger F-statistic is needed to reject the null hypothesis

Multiple Linear Regression

1. Is there a relationship between the response and predictors?

How large does the F-statistic need to be before we can reject the null hypothesis and conclude that there is a relationship?

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)} \sim F_{p,n-p-1} \quad \text{if} \quad \begin{cases} \text{The errors have a normal distribution} \\ \text{The sample size n is sufficiently large} \end{cases}$$

We can compute the p-value associated with the F-statistic using the F-distribution and then determine whether or not to reject the null hypothesis.

Multiple Linear Regression

1. Is there a relationship between the response and a subset of predictors?

Testing the *null hypothesis* of

$$H_0 : \beta_{p-q+1} = \beta_{p-q+2} = \ldots = \beta_p = 0$$

$$F = \frac{(RSS_0 - RSS)/q}{RSS/(n-p-1)}$$

RSS_0 : the residual sum of squares obtained from fitting a second model of using all the variables except those last q.

Multiple Linear Regression

1. Is there a relationship between the response and a subset of predictors?

Advertising data

	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

→ Provides information on whether each individual predictor is related to the response, after adjusting for the other predictors

‖

The F-test that omits that single variable, leaving all the others in

Multiple Linear Regression

1. Is there a relationship between the response and predictors?

Given these individual p-values, why the overall F-statistic?

> “If any one of the p-values for the individual variables is very small,
> then at least one of the predictors is related to the response” ✗

Multiple Linear Regression

1. Is there a relationship between the response and predictors?

Example: p=100, $H_0 : \beta_1 = \beta_2 = \ldots = \beta_{100} = 0$

At a level of 0.05, we are allowing a 5% chance of incorrectly rejecting the null hypothesis. When p=100, we expect to see approximately 5 small p-values even in the absence of any true association between the predictors and the response.

The F-statistic does not suffer from this problem. If the null hypothesis is true, there is only a 5% chance that the F-statistic will result in a p-value below 0.05, regardless of the number of predictors or the number of observations.