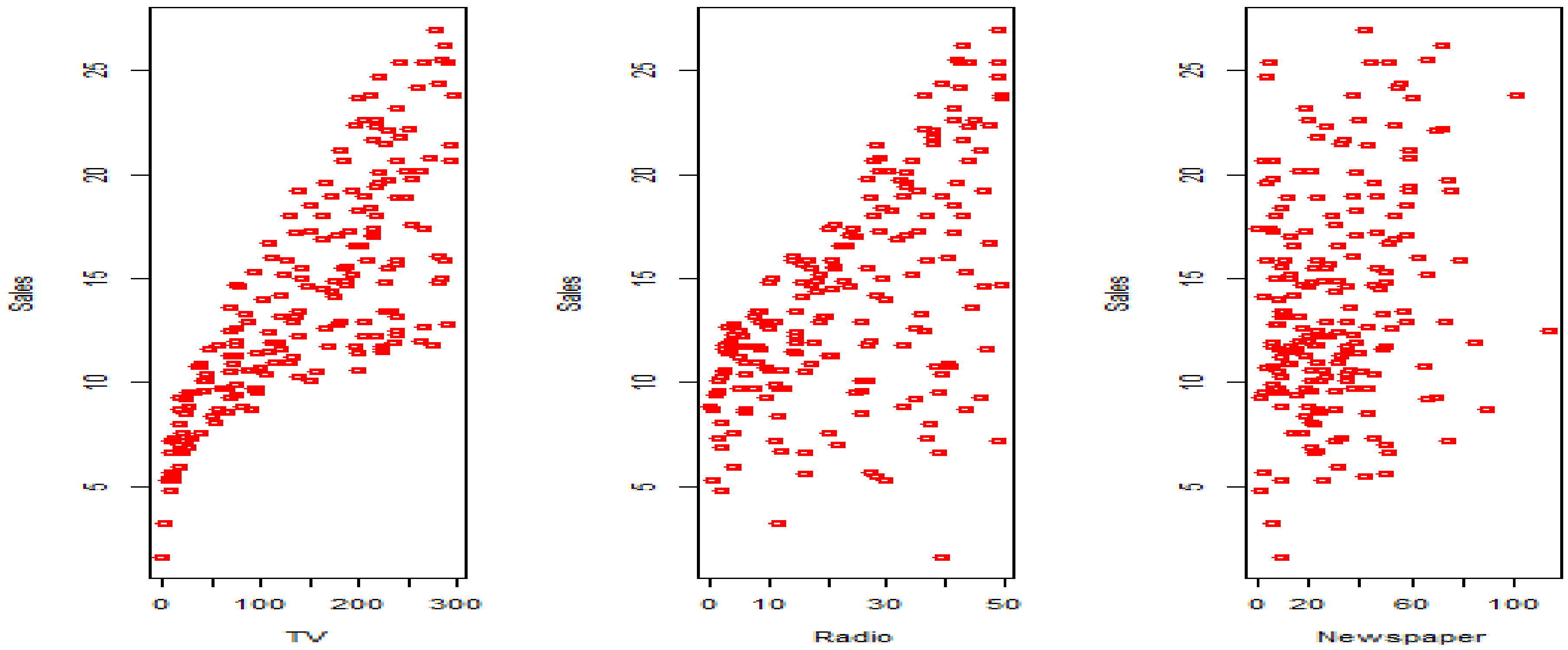
Simple Linear Regression

Advertising Data Set

- The *Advertising* data set consists of the *sales* (in thousands of units) of a particular product in 200 different markets.
- It also contains the advertising budgets (in thousands of dollars) for the product in each of the markets for three different media: *TV*, *radio*, and *newspaper*.
- Our objective is to check the association between *advertising budgets* and sales.

Plot of *Advertising* Data Set



Notation

- Output Variable/ Response Variable/ Dependent Variable: *Sales* (Y).
- Input Variables/ Predictors/ Independent Variables/ Features/ Variables:
  - *TV budget* (X_1)
  - *Radio budget* (X_2)
  - *Newspaper budget* (X_3)

Important Questions for an Effective Market Plan

1. Is there any relationship between advertising budget and sales?
2. How strong is the relationship between advertising budget and sales?
3. Which media contribute to sales?
4. How accurately can we estimate the effect of each medium on sales?
5. How accurately can we predict the future sales?
6. Is the relationship linear?
7. Is there synergy among the advertising media?

Simple Linear Regression (SLR)

- Y: Response
- X: Predictor
- In SLR, we assume that there is approximately a linear relationship between X and Y.
- Thus the SLR model is given by

$$Y = \beta_0 + \beta_1 X + \epsilon, \ldots \ldots \ldots (1)$$

where ϵ is mean-zero random error term.

- For example, let X represent *TV* advertising budget and Y represent *sales*, then the SLR model is given by

$$sales = \beta_0 + \beta_1 \times TV + \epsilon.$$

Prediction

- In Equation (1), β_0 and β_1 are two unknown constants that represent the intercept and slope terms in the linear model.
- Together they are called the *model coefficients or parameters*.
- Let the estimates of β_0 and β_1 based on the training data be $\hat{\beta}_0$ and $\hat{\beta}_1$.
- Now, given the budget of *TV* advertising, we can predict future sales by computing

$$\widehat{sales} = \hat{\beta}_0 + \hat{\beta}_1 \times TV,$$

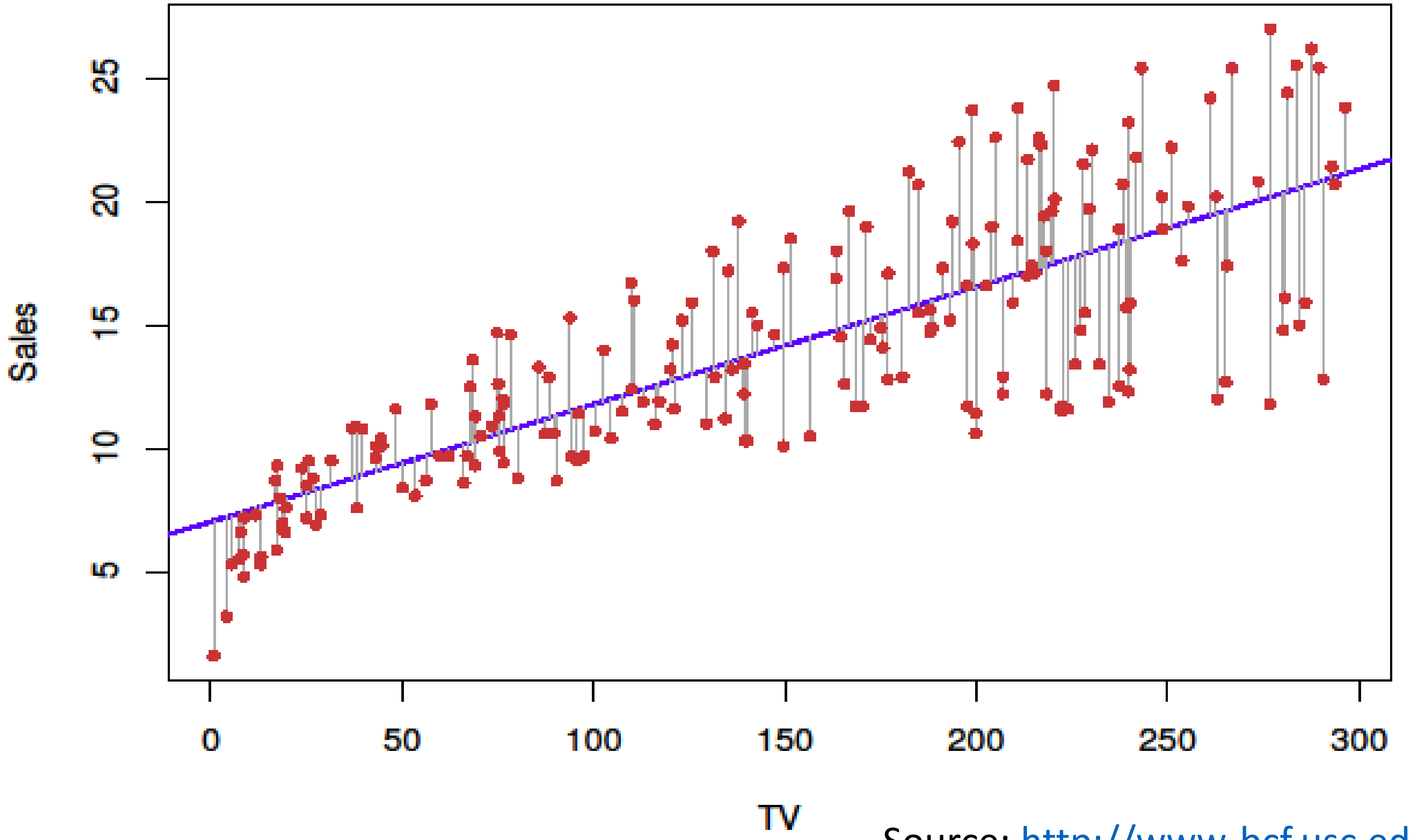
where

$\widehat{sales}$: predicted sales.

How to Get the Estimates of the Coefficients?

- In the *advertising* data set, we have data on the TV advertising budget and product sales in $n = 200$ different markets.
- Our job is to find the coefficients in such a way that the resulting line is as close as possible to the $n = 200$ data points.
- How should we measure the *closeness*?
- The most common approach involves minimizing *the least squares criterion*.

How to Get the Estimates of the Coefficients?



Source: http://www-bcf.usc.edu/~gareth/ISL/

How to Get the Estimates of the Coefficients?

- Data: $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$.
- Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for Y based on the i-th value of X.
- Residual: $e_i = y_i - \hat{y}_i$.
- We define the *residual sum of squares* (RSS) as

$$RSS = e_1^2 + \cdots + e_n^2.$$

- We minimize RSS to get the estimates of the coefficients.

Summary of Regression Analysis of *Sales* on *TV*

	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

Interpretations

- An increase of \$1000 in the TV advertising budget is associated with an increase in average sales by around 48 units.
- Accuracy of the estimated coefficients can be measured from their respective standard errors.
- Standard errors can be used to construct the *confidence intervals*.
- A 95% confidence interval is defined as the range of values such that with 95% probability, the range will contain the true unknown value of the parameter.

Interpretations

- For linear regression, the 95% confidence interval for β_1 approximately takes the form
$$\hat{\beta}_1 \pm 2.SE(\hat{\beta}_1).$$
- That is, there is approximately a 95% chance that the interval
$$\left[\hat{\beta}_1 - 2.SE(\hat{\beta}_1), \hat{\beta}_1 + 2.SE(\hat{\beta}_1)\right]$$
will contain the true value of β_1.
- Similarly, confidence interval for β_0 approximately takes the form
$$\hat{\beta}_0 \pm 2.SE(\hat{\beta}_0).$$

Interpretations

- In the case of the advertising data, the 95% confidence interval for β_0 is [6.130,7.935] and the 95% confidence interval for β_1 is [0.042,0.053].
- Therefore, we can conclude that in the absence of any advertising, sales will, on average, fall somewhere between 6,130 and 7,940 units.
- Furthermore, for each $1,000 increase in television advertising, there will be an average increase in sales of between 42 and 53 units.

Interpretations

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

$$H_0 : There\ is\ no\ relationship\ between\ X\ and\ Y$$

versus the alternative

$$H_1 : There\ is\ some\ relationship\ between\ X\ and\ Y.$$

- Mathematically, this corresponds to testing

$$H_0 : \beta_1 = 0\ vs\ H_1 : \beta_1 \neq 0.$$

Interpretations

- The test statistic for testing the hypothesis is given by
$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}.$$
- Under H_0, the above test statistic follows a t-distribution with $n - 2$ degrees of freedom.
- We reject the null hypothesis, i.e., we declare that no relationship exists between X and Y, if the p-value is small enough.

Interpretations

- Notice that the coefficients are very large relative to their standard errors, so the t-statistics are also large.
- The probabilities of seeing such values if H_0 is true are virtually zero. Hence we can conclude that $\beta_0 \neq 0$ and $\beta_1 \neq 0$.
- This clearly suggests that there is a significant relationship between *TV* and *sales*.

Assessing the Accuracy of the Model

- Once the null hypothesis is rejected in favour of the alternative hypothesis, it is natural to quantify the extent to which the model fits the data.
- The quality of a linear regression fit is assessed using two related quantities:
  - Residual Standard Error
  - R^2 statistic

Residual Standard Error (RSE)

- The RSE is an estimate of the standard deviation of ϵ.
- Roughly, it is the average amount by which the response will deviate from the true regression line.
- It is computed using the formula

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

- The RSE is considered a measure of the *lack of fit*.

Residual Standard Error (RSE)

- In case of the advertising data, the RSE is 3.26.
- This means that the 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 β_0 and β_1 were known exactly, any prediction of sales on the basis of TV advertising would still be off by about 3,260 units on average.
- The next question is whether or not 3,260 units is an acceptable prediction error.

Residual Standard Error (RSE)

- In the advertising data set, the mean value of sales over all markets is approximately 14,000 units, and so the percentage error is 3,260/14,000 = 23%.

R^2 Statistic

- The RSE provides an absolute measure of lack of fit of the model to the data.
- 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 .

R^2 Statistic

- To calculate R^2, we use the formula
$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS},$$
where $TSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is the *total sum of squares*.
- TSS measures the amount of variability inherent in the response before the regression is performed.
- In contrast, RSS measures the amount of variability that is left unexplained after performing the regression.
- Hence, $TSS - RSS$ 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*.

R^2 Statistic

- An R^2 statistic 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.
- This might occur because the linear model is wrong, or the inherent error σ^2 is high, or both.
- In the advertising data set, the R^2 was 0.61, and so just under two-thirds of the variability in sales is explained by a linear regression on TV alone.

R^2 Statistic

- The R^2 statistic has an interpretational advantage over the *RSE*, since unlike the RSE, it always lies between 0 and 1.
- However, it can still be challenging to determine what is a good R^2 value, and in general, this will depend on the application.
- In the simple linear regression setting, $R^2 = r^2$.
- Thus R^2 can work as a measure of linear relationship between X and Y.

Summary of Regression Analysis 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

Summary of Regression Analysis 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	0.0012

Summary of Three Simple Linear Regression Models

Model	Predictors	R^2	RSE
1	TV	0.62	3.26
2	Radio	0.33	4.28
3	Newspaper	0.05	5.09

Fitted Regression Lines