LINEAR REGRESSION

Part I

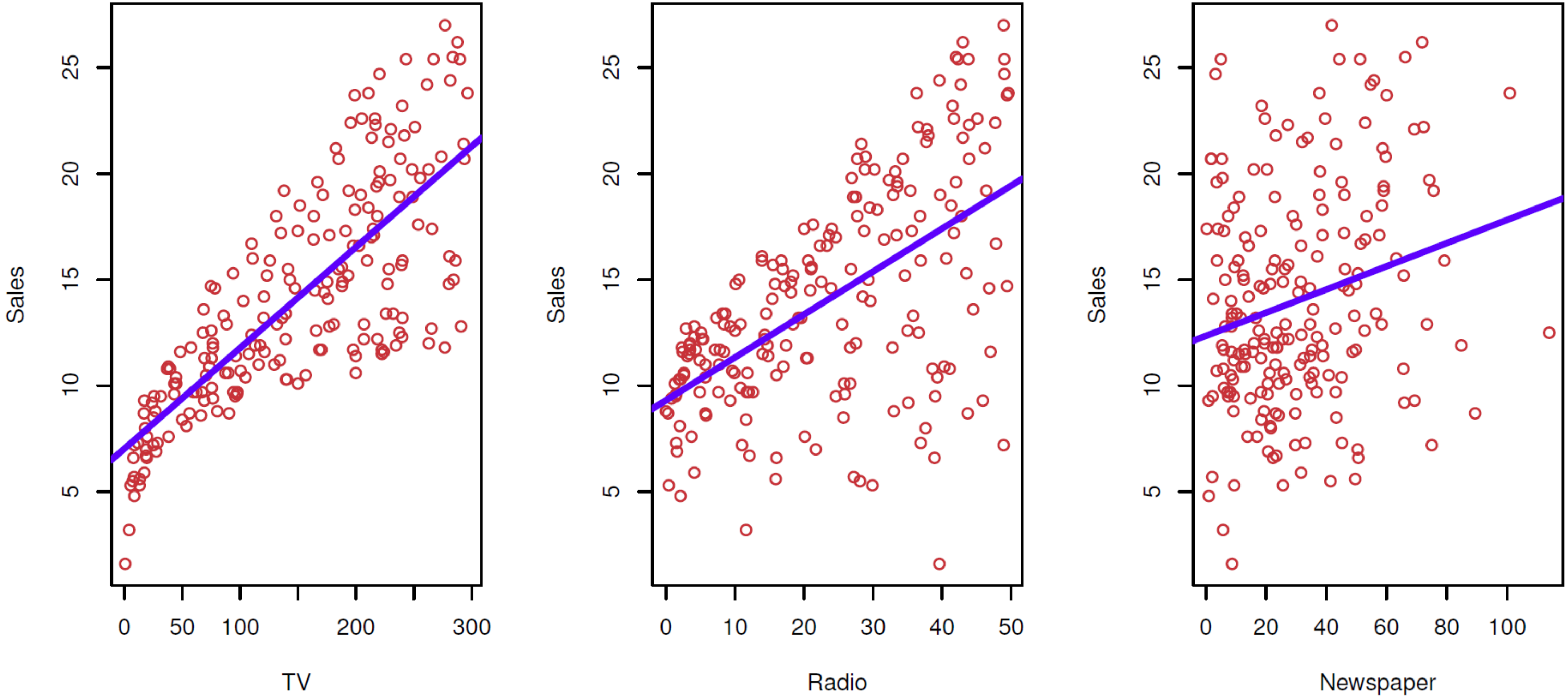
Outline

- Simple linear regression

Advertising data

- Provide advice on how to improve sales of a product
- The Advertising data set consists of the sales of that product in 200 different markets, along with advertising budgets for three different media: TV, radio, and newspaper

- *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 the relationship linear?*
- *Is there synergy among the advertising media?*



Some of the figures and tables in this presentation are taken from "*An Introduction to Statistical Learning, with Applications in R*" (Springer) with permission from the authors: G. James, D. Witten, T. Hastie, and R. Tibshirani

Simple linear regression

- The simple linear regression model

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

  - β_0 and β_1 are unknown parameters or coefficients
- Explanation?

- β_0 is the expected value of Y when $X = 0$
- β_1 represents the average increase in Y associated with a one-unit increase X
- ϵ is the error term

- Apply a statistical learning method to the training data to obtain coefficient estimates $\hat{\beta}_0$ and $\hat{\beta}_1$
- Prediction formula

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Estimating the coefficients

- Let $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ represent n observation pairs
- Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ and $e_i = y_i - \hat{y}_i$
  - e_i's are known as residuals

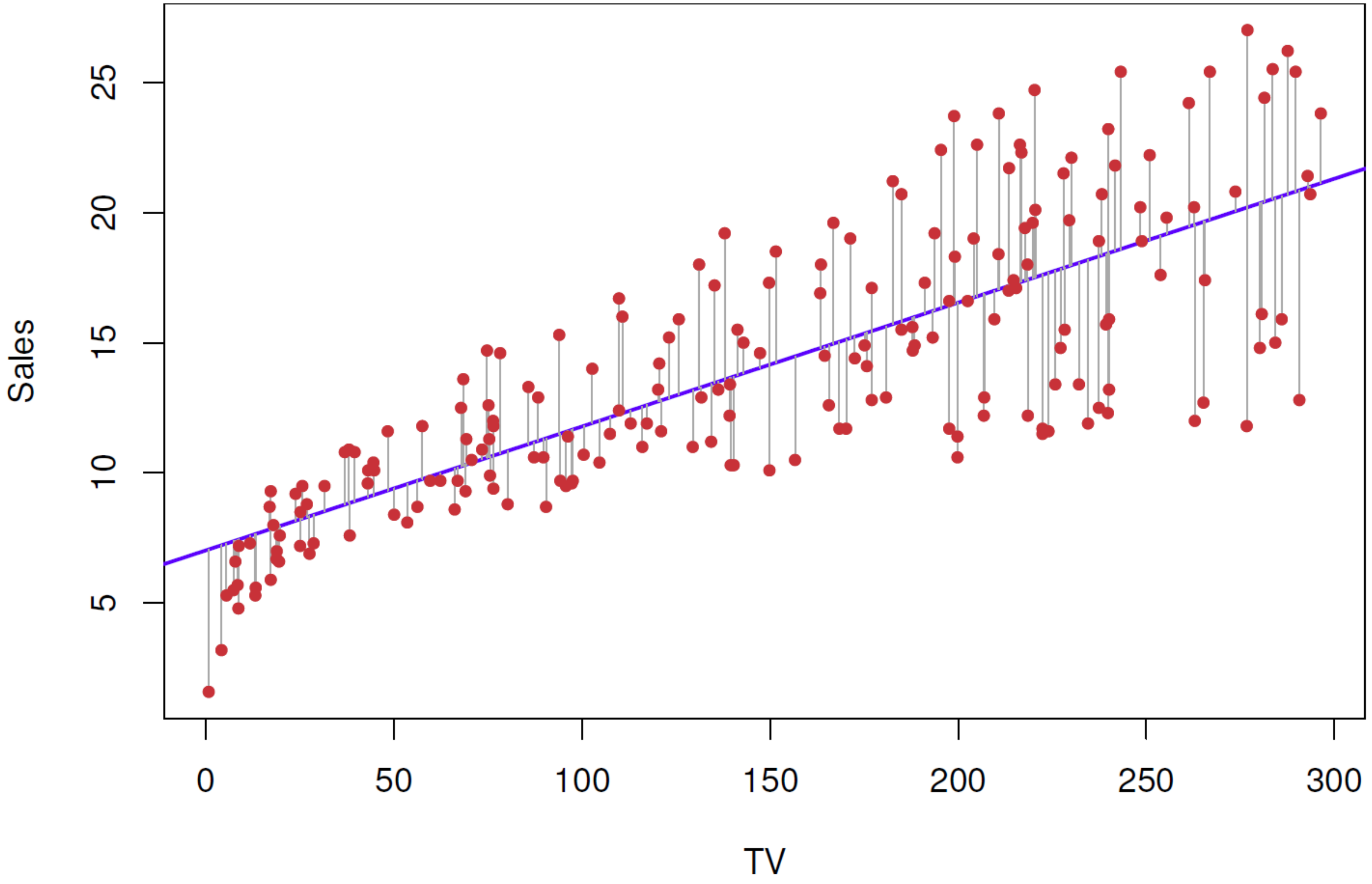
- Define the residual sum of squares

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

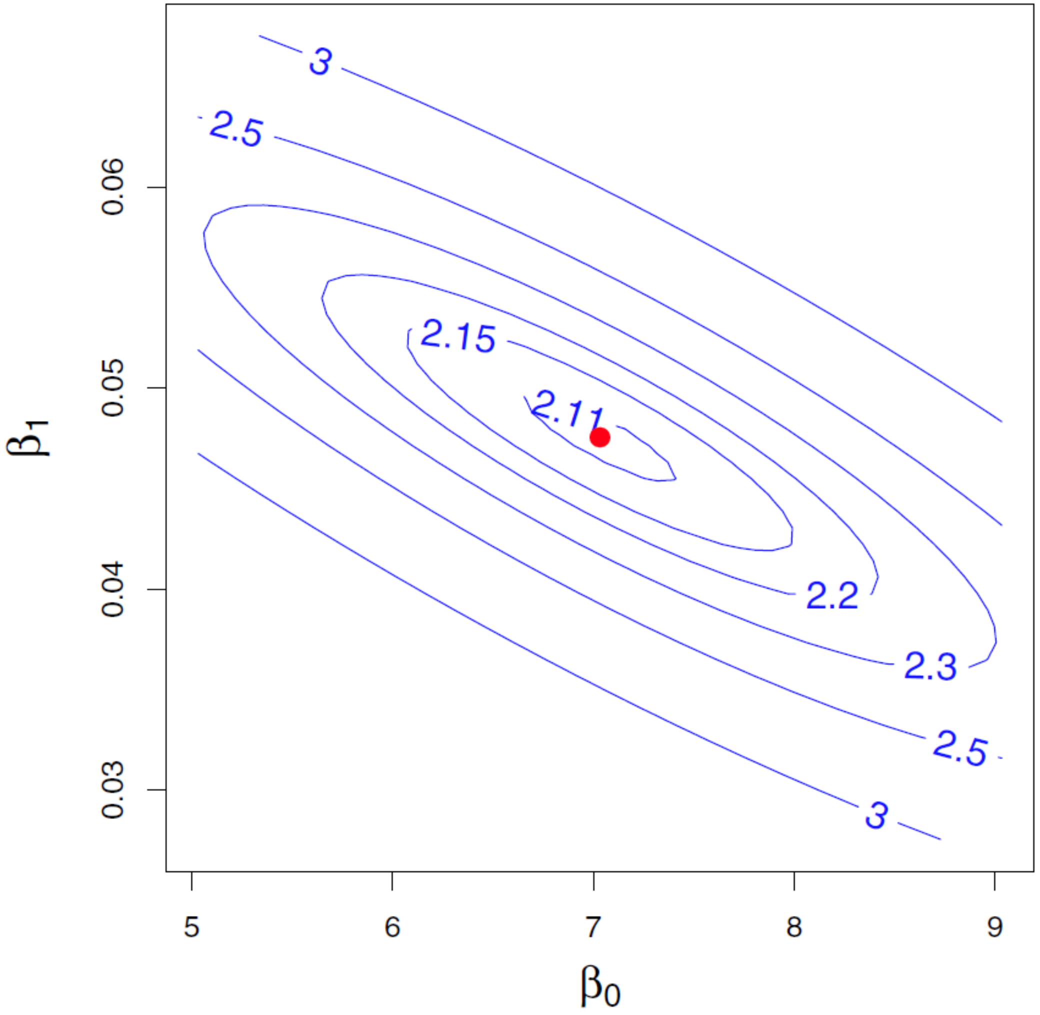
- The least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS

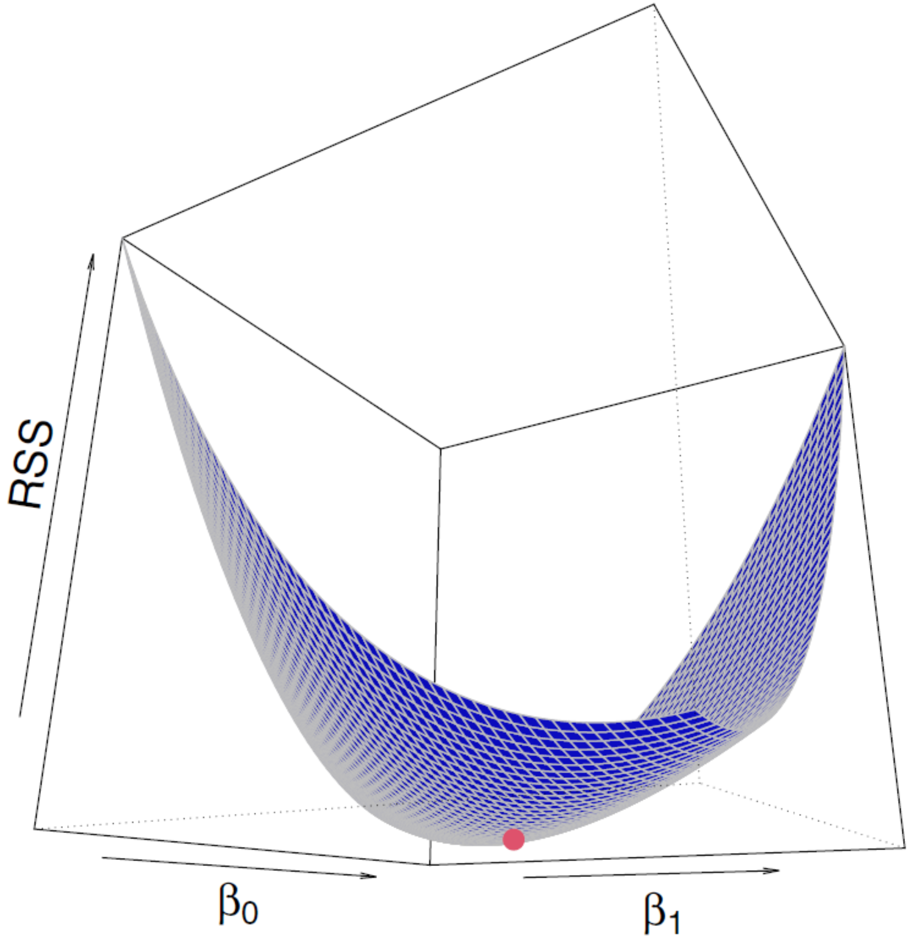
Least squares coefficient estimates

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



The least squares fit, $\hat{y} = 7.03 + 0.0475x$, for the regression of sales onto TV





Contour and three-dimensional plots of the RSS on the Advertising data, using sales as the response and TV as the predictor

Properties of the coefficient estimates

- The population regression line

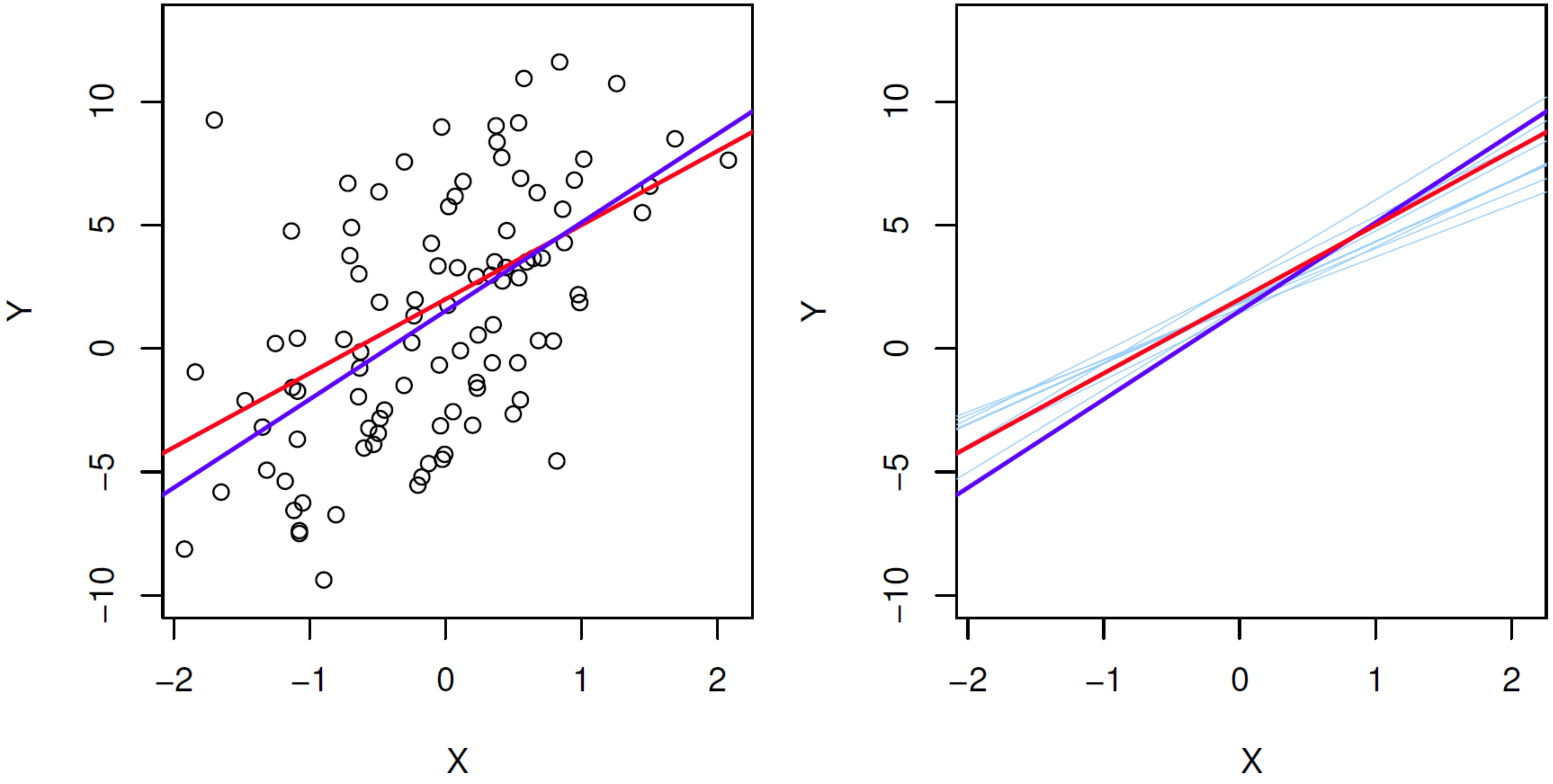
$$E(Y) = f(X) = \beta_0 + \beta_1 X$$

- The least squares line

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

- The least squares coefficient estimates are *unbiased*

$$E(\hat{\beta}_0) = \beta_0$$
$$E(\hat{\beta}_1) = \beta_1$$



A simulated example with $Y = 2 + 3X + \epsilon$

- Variances

$$\mathrm{Var}\left(\hat{\beta}_0\right) = \sigma^2 \left\{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right\}$$

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

- Residual standard error (RSE)

$$\mathrm{RSE} = \hat{\sigma} = \sqrt{\frac{\mathrm{RSS}}{n-2}}$$

- Standard errors

$$\mathrm{SE}\left(\hat{\beta}_0\right) = \sqrt{\hat{\sigma}^2\left\{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right\}}$$

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

Confidence intervals

- A $(1-\alpha)100\%$ confidence interval is defined as a range of values such that with $(1-\alpha)100\%$ probability, the range will contain the true unknown value of the parameter

- The 95% confidence interval for β_0 *approximately* takes the form

$$\hat{\beta}_0 \pm 2 \times \mathrm{SE}\left(\hat{\beta}_0\right)$$

- The 95% confidence interval for β_1 *approximately* takes the form

$$\hat{\beta}_1 \pm 2 \times \mathrm{SE}(\hat{\beta}_1)$$

- For the Advertising data, the 95% confidence interval for β_0 is $[6.130, 7.935]$, and the 95% confidence interval for β_1 is $[0.042, 0.053]$

Hypothesis tests

- The null hypothesis

  H_0: There is no relationship between X and Y or, $\beta_1 = 0$

- The alternative hypothesis

  H_a: There is some relationship between X and Y or, $\beta_1 \neq 0$

t-test

- Testing H_0: $\beta_1 = 0$ versus H_a: $\beta_1 \neq 0$
- t-statistic

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

- Under H_0: $\beta_1 = 0$, t has a t-distribution with $n - 2$ degrees of freedom

p-value

- The probability of observing any value equal to $|t|$ or larger, assuming $\beta_1 = 0$
- In the absence of any real association, a small p-value indicates that it is unlikely to observe such a substantial association due to chance

- 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

TABLE 3.1. *For the* `Advertising` *data, coefficients of the least squares model for the regression of number of units sold on TV advertising budget. An increase of* \$1,000 *in the TV advertising budget is associated with an increase in sales by around 50 units. (Recall that the* `sales` *variable is in thousands of units, and the* `TV` *variable is in thousands of dollars.)*

Quality of the least squares fit

- Two measures of the *lack of fit*
  - Residual standard error
  - R^2 statistic

- Residual standard error

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

Quantity	Value
Residual standard error	3.26
R^2	0.612
F-statistic	312.1

TABLE 3.2. *For the* `Advertising` *data, more information about the least squares model for the regression of number of units sold on TV advertising budget.*

- R^2 statistic

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

  - $\mathrm{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
  - $\mathrm{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$
- R^2 measures the *proportion of variability* in Y that can be explained using X

- Exercise: Show that $R^2 = r^2$, where

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Quantity	Value
Residual standard error	3.26
R^2	0.612
F-statistic	312.1

TABLE 3.2. *For the* `Advertising` *data, more information about the least squares model for the regression of number of units sold on TV advertising budget.*