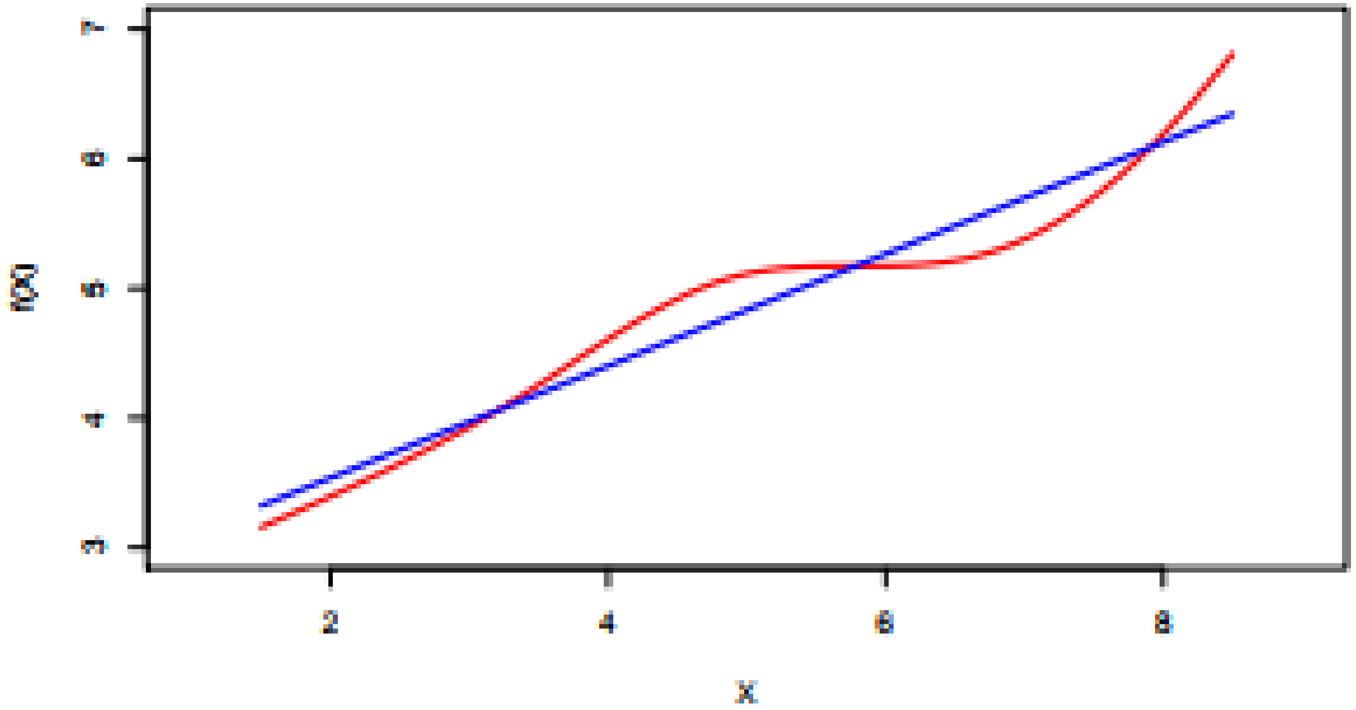
Linear Regression

Linear regression

- Linear regression is a simple approach to supervised learning.
- It assumes that the dependence of Y on $X_1, X_2, \ldots X_p$ is linear.



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

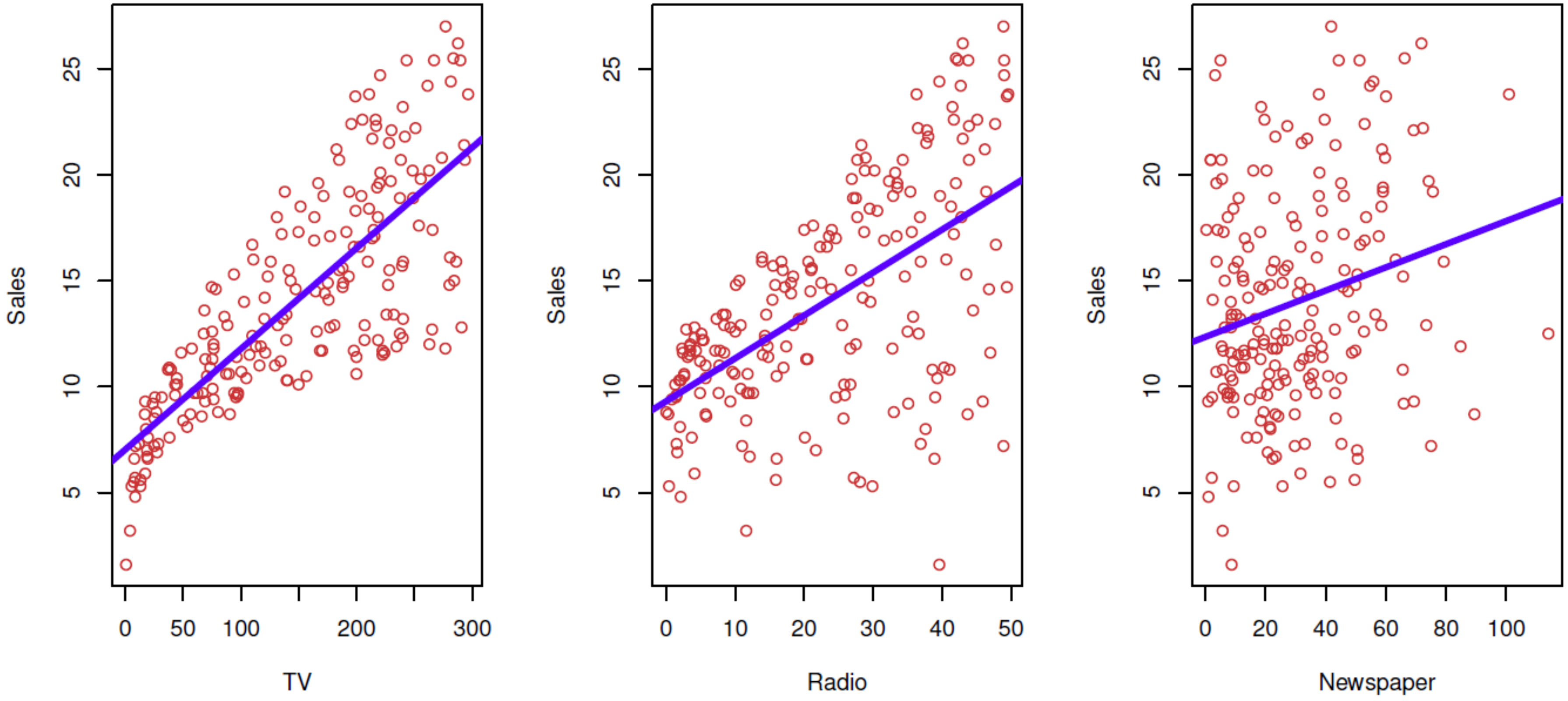
Linear regression for the advertising data

- Consider the advertising data shown on the next slide.

Questions we might ask:

  - 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 predict future sales?
  - Is the relationship linear?
  - Is there synergy among the advertising media?

Advertising data



Simple linear regression using a single predictor X

- We assume a model

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

- where β_0 and β_1 are two unknown constants that represent the *intercept* and *slope*, also known as *coefficients* or *parameters*, and is the error term.
- Given some estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for the model coefficients, we predict future sales using

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

- where $\hat{y}$ indicates a prediction of Y on the basis of $X = x$. The hat symbol denotes an estimated value.

Estimation of the parameters by least squares

- 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. Then $e_i = y_i - \hat{y}_i$ represents the *i*th *residual*
- We define the *residual sum of squares* (RSS) as

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

or equivalently as

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

- The least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS. The minimizing values can be shown to be

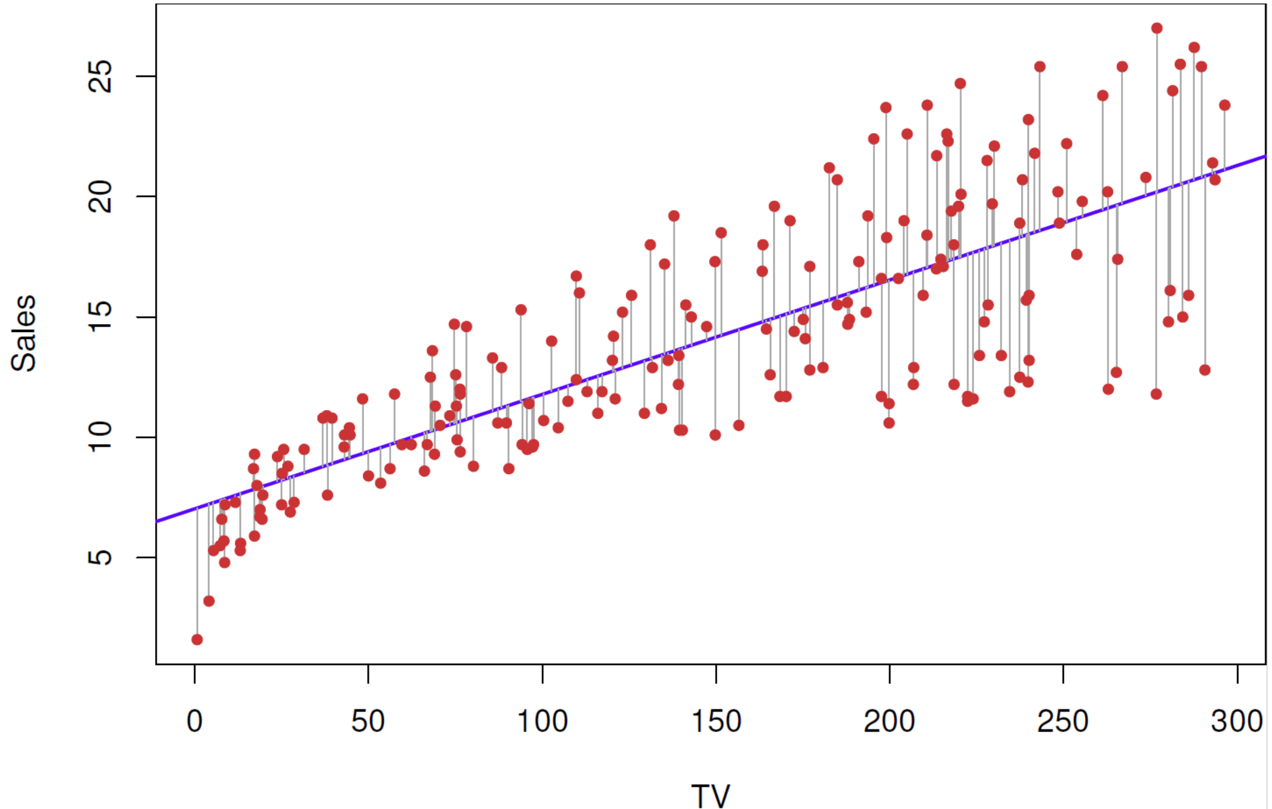
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

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

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

Example: advertising data

- The least squares fit for the regression of sales onto TV. In this case a linear fit captures the essence of the relationship, although it is somewhat deficient in the left of the plot



Assessing the Accuracy of the Coefficient Estimates

- The standard error of an estimator reflects how it varies under repeated sampling. We have

$$\mathrm{SE}(\hat{\beta}_1{}^2) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \mathrm{SE}(\hat{\beta}_0{}^2) = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$$

where $\sigma^2 = \mathrm{Var}(\epsilon)$

- These standard errors can be used to compute confidence intervals. A 95% confidence interval is defined as a range of values such that with 95% probability, the range will contain the true unknown value of the parameter. It has the form $\hat{\beta}_1 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_1)$.

Confidence intervals — continued

- That is, there is approximately a 95% chance that the interval

$$[\hat{\beta}_1 - 2 \cdot \mathrm{SE}(\hat{\beta}_1), \hat{\beta}_1 + 2 \cdot \mathrm{SE}(\hat{\beta}_1)]$$

will contain the true value of β_1 (under a scenario where we got repeated samples like the present sample)

- For the advertising data, the 95% confidence interval for β_1 is

[0.042, 0.053]

Hypothesis testing

- 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 hypothesis*

  H_A : There is some relationship between X and Y.

- Mathematically, this corresponds to testing

  $H_0 : \beta_1 = 0$

  versus

  $H_A : \beta_1 \neq 0$

since if $\beta_1 = 0$, then the model reduces to $Y = \beta_0 + \epsilon$, and X is not associated with Y.

Hypothesis testing — continued

- To test the null hypothesis, we compute a *t-statistic*, given by

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

- This will have a t-distribution with n−2 degrees of freedom, assuming $\beta_1 = 0$.
- Using statistical software, it is easy to compute the probability of observing any value equal to |t| or larger. We call this probability the *p-value*.

Results for the advertising data

	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

Assessing the Overall Accuracy of the Model

- We compute the *Residual Standard Error*

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

where the residual *sum-of-squares* is $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

- *R-squared* or fraction of variance explained is

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

where $\mathrm{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$

is the *total sum of squares*.

- It can be shown that in this simple linear regression setting that

$R^2 = r^2$, where r is the correlation between X and Y

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

Advertising data results

Quantity	Value
Residual Standard Error	3.26
R^2	0.612
F-statistic	312.1

Multiple Linear Regression

- Here our model is $Y = \beta 0 + \beta 1X1 + \beta 2X2 +\cdots+ \beta pXp +$, • We interpret βj as the average effect on Y of a one unit increase in Xj, holding all other predictors fixed. In the advertising example, the model becomes
- $sales = \beta 0 + \beta 1 \times TV + \beta 2 \times radio + \beta 3 \times newspaper +$.

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 Xj is associated with a βj change in Y , while all the other variables stay fixed”, are possible. • Correlations amongst predictors cause problems: - The variance of all coefficients tends to increase, sometimes dramatically - Interpretations become hazardous — when Xj changes, everything else changes. • Claims of causality should be avoided for observational data.

The woes of (interpreting) regression coefficients

- “Data Analysis and Regression” Mosteller and Tukey 1977 • a regression coefficient βj estimates the expected change in Y per unit change in Xj, with all other predictors held fixed. But predictors usually change together!
- Example: Y total amount of change in your pocket; X1 = # of coins; X2 = # of pennies, nickels and dimes. By itself, regression coefficient of Y on X2 will be > 0. But how about with X1 in model?
- Y = number of tackles by a football player in a season; W and H are his weight and height. Fitted regression model is ˆ Y = b0 + .50W −.10H. How do we interpret ˆ β2 < 0?

Two quotes by famous Statisticians

- “Essentially, all models are wrong, but some are useful”
- 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”
- Fred Mosteller and John Tukey, paraphrasing George Box

Estimation and Prediction for Multiple Regression

- Given estimates $\hat{\beta}_0$, $\hat{\beta}_1$,... $\hat{\beta}_p$, we can make predictions using the formula
- $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$. • We estimate $\beta_0, \beta_1, \ldots, \beta_p$ as the values that minimize the sum of squared residuals
- RSS =
- $\sum_{i=1}^{n}$
- $(y_i - \hat{y}_i)^2$
- =
- $\sum_{i=1}^{n}$
- $(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2$.
- This is done using standard statistical software. The values $\hat{\beta}_0$, $\hat{\beta}_1$,..., $\hat{\beta}_p$ that minimize RSS are the multiple least squares regression coefficient estimates.

- Image 19/48

Results for advertising data

Some important questions

- 1. Is at least one of the predictors X1,X2,...,Xp useful in predicting the response?
- 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?

Is at least one predictor useful?

- For the first question, we can use the F-statistic
- F =
- (TSS−RSS)/p RSS/(n−p−1) ∼ Fp,n−p−1
- Quantity Value Residual Standard Error 1.69 R2 0.897 F-statistic 570

Deciding on the important variables

- The most direct approach is called all subsets or best subsets regression: we compute the least squares fit for all possible subsets and then choose between them based on some criterion that balances training error with model size.
- However we often can't examine all possible models, since they are 2p of them; for example when $p = 40$ there are over a billion models! Instead we need an automated approach that searches through a subset of them. We discuss two commonly use approaches next.

Forward selection

- Begin with the null model — a model that contains an intercept but no predictors. • Fit p simple linear regressions and add to the null model the variable that results in the lowest RSS. • Add to that model the variable that results in the lowest RSS amongst all two-variable models. • Continue until some stopping rule is satisfied, for example when all remaining variables have a p-value above some threshold.

Backward selection

- Start with all variables in the model. • 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. • Continue until a stopping rule is reached. For instance, we may stop when all remaining variables have a significant p-value defined by some significance threshold.

Model selection — continued

- Later we discuss more systematic criteria for choosing an "optimal" member in the path of models produced by forward or backward stepwise selection.

• These include Mallow's Cp, Akaike information criterion (AIC), Bayesian information criterion (BIC), adjusted R2 and Cross-validation (CV).