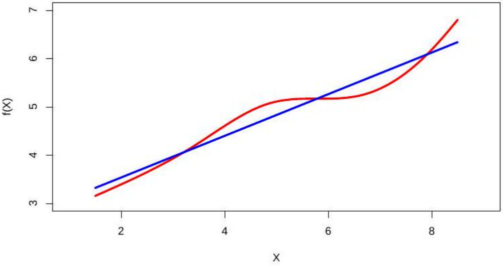
Data Mining

19th February 2023

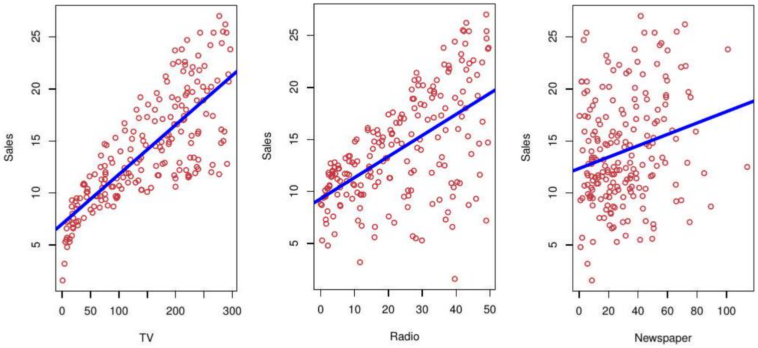
Linear Regression ***

- 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.
- True regression functions are never linear!



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

Advertising data



Linear regression for the advertising data

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?

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 ϵ 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.

Simple linear regression using a single predictor X.

- Usually, we assume that $\text{cov}(X, \epsilon) = 0$ or, more strictly, ϵ is independent of X.
- The assumption ϵ is independent of X makes it easier to derive the maximum likelihood estimation (MLE).
- Consider n i.i.d samples $\{(X_i, Y_i)\}_{i=1}^n$, we usually assmue that $\text{var}(\epsilon_i) = \sigma^2$ for all i.

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

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

or equivalently as

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

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

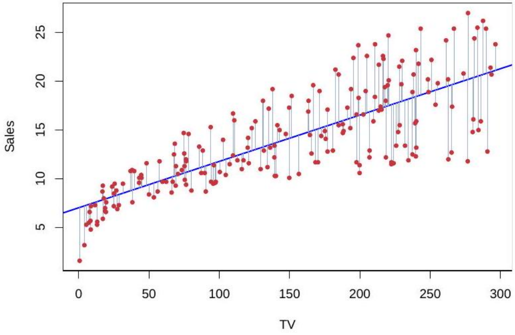
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2},$$

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

Estimation of the parameters by maximum likelihood

- Suppose that $y_i = \beta_0 + \beta_1 x_i$, for $i = 1, \cdots, n$. We further assume that $\epsilon_i \sim N(0, \sigma^2)$ and ϵ_i is independent of x_i, for all $i = 1, \cdots, n$.
- Question: what are the maximum likelihood estimations (MLEs) of β_0 and β_1?
- Will the MLEs change if $\epsilon_i \sim N(0, \sigma_i^2)$?

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

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

  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

$$\hat{\beta}_1 \pm 2 \cdot \mathrm{SE}\left(\hat{\beta}_1\right).$$

Confidence intervals - continued

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

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

  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 $[0.042, 0.053]$

Confidence intervals - continued

- Note that the variance σ^2 is usually unknown in practice.
- We estimate it by $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n-2}$.
- Why we use $n-2$?

Hypothesis testing

- Next, we consider the hyootheis test for β_1. Here we assume that $\epsilon_i \sim N(0, \sigma^2)$.
- 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

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

- 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

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

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

- R-squared or fraction of variance explained is

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

where $\mathsf{TSS} = \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2$ is the total sum of squares.

Assessing the Overall Accuracy of the Model

- 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 :

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

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

Multiple Linear Regression

- Here our model is

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

- We interpret β_j as the average effect on Y of a one unit increase in X_j, holding all other predictors fixed. In the advertising example, the model becomes

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

Estimation and Prediction for Multiple Regression

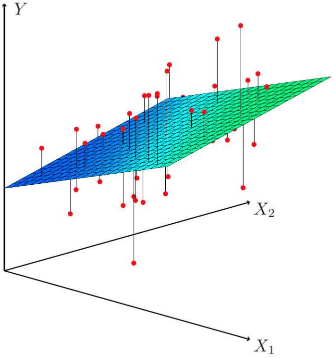
- Given estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots \hat{\beta}_p$, we can make predictions using the formula

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

- We estimate $\beta_0, \beta_1, \ldots, \beta_p$ as the values that minimize the sum of squared residuals

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

This is done using standard statistical software. The values $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ that minimize RSS are the multiple least squares regression coefficient estimates.



Results for 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

	Correlations:			
	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

Some important questions

1. Is at least one of the predictors $X_1, X_2, \ldots, X_p$ 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?

- For the first question, we can use the F-statistic

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

Quantity	Value
Residual Standard Error	1.69
R^2	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 2^p 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 C_p, Akaike information criterion (AIC), Bayesian information criterion (BIC), adjusted R^2, Cross-validation (CV), (Ridge), Lasso, Elastic net, Dimension reduction approaches and Sure independence screening (SIS).

Extensions of the Linear Model

- Multivariate linear regression, reduced rank regression and canonical correlation analysis.
- Categorical response and logistic regression.
- High-dimensional problem: $p \gg n$.