Statistical Learning and Data Mining

Module 2: Linear Regression and Statistical Thinking

Semester 2, 2017

Discipline of Business Analytics, The University of Sydney Business School

Module 2: Linear Regression and Statistical Thinking

1. Introduction
2. The least squares algorithm
3. The MLR model
4. Statistical properties
5. The Gaussian MLR model
6. Interpreting a linear regression model

Introduction

Linear regression

The linear regression is a simple and widely used method for supervised learning. There are several important reasons for developing an in-depth understanding of this method.

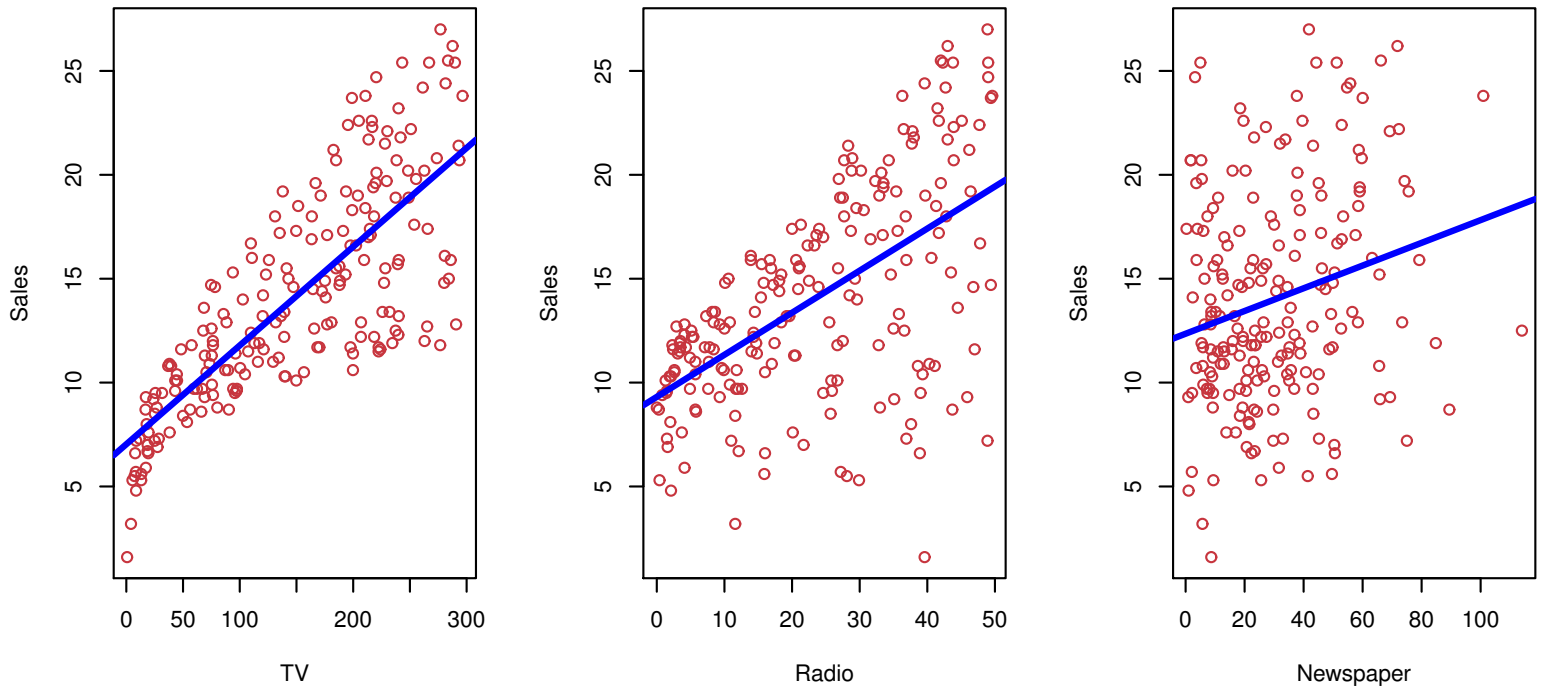
- It is very useful for prediction in many settings.
- It is extremely useful conceptually. Many advanced statistical learning methods can be understood as extensions and generalisations of linear regression.
- Due to its simplicity, linear regression is often a useful jumping-off point for model building and analysis.
- Interpretability.

Example: advertisement data

Consider from example the advertisement data from the ISL textbook (see next slide). Possible questions:

- 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 there synergy among the advertisement media?

Example: advertisement data



(Figure from ISL)

Example: advertisement data

To answer our questions, we can use a model such as

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

Statistical thinking

Statistical thinking is using statistical models, statistical theory, and critical thinking to learn from data.

- How do I design a study to answer a certain question?
- How relevant and representative are my data?
- What is the variability in my data? Can I reliably draw conclusions in light of this variability?
- How do I correctly interpret my results?
- Can I generalise my conclusions in the way that I would like to?
- What are the limitations of my analysis?

The least squares algorithm

Linear regression

In the linear regression method for prediction, we consider a regression function of the form

$$f(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

We learn the prediction coefficients $\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p$ by fitting the model to the training data using the least squares method.

Least squares

Let $\mathcal{D} = \{(y_i, \boldsymbol{x}_i)\}_{i=1}^N$ be the training data. We define the **residual sum of squares** as a function of parameter values $\boldsymbol{\beta}$ as

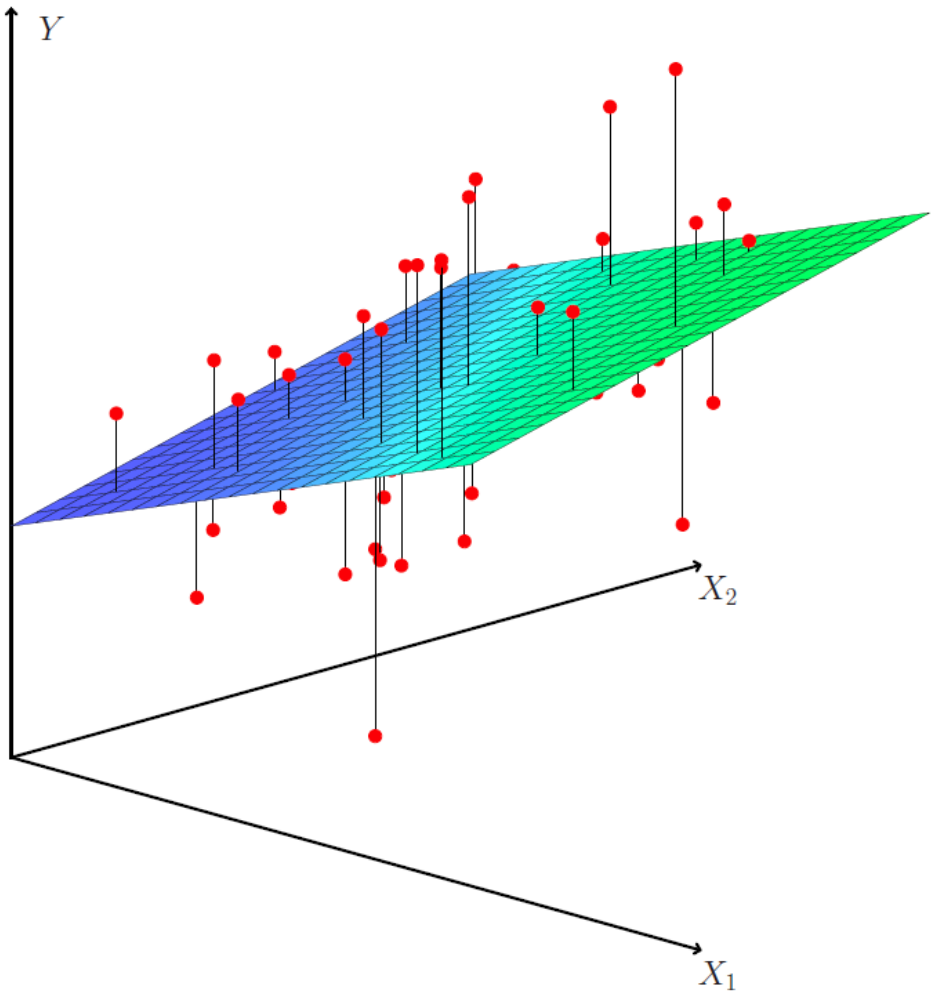
$$\begin{aligned}\mathsf{RSS}(\boldsymbol{\beta}) &= \sum_{i=1}^N (y_i - f(\boldsymbol{x}_i;\, \boldsymbol{\beta}))^2 \\ &= \sum_{i=1}^N \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2\end{aligned}$$

Least squares

The **ordinary least squares** (OLS) method selects the coefficient values that minimise the residual sum of squares

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{argmin}} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

Least squares



(Figure from ISL)

Interpretation

If our loss function $L(y, f(\boldsymbol{x}))$ is the squared error loss, the OLS algorithm consists of minimising the empirical loss for our choice of predictive function:

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{argmin}} \ \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(\boldsymbol{x}_i))$$

Least squares and linear algebra

In order to obtain a solution to the OLS minimisation problem, we need linear algebra.

https://xkcd.com/1838/

Design matrix

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & x_{N2} & \dots & x_{Np} \end{pmatrix}$$

Least squares and linear algebra

We equivalently write the RSS as

$$\begin{aligned}\text{RSS}(\beta) &= (\boldsymbol{y} - \boldsymbol{X}\beta)^T(\boldsymbol{y} - \boldsymbol{X}\beta) \\ &= \boldsymbol{y}^T\boldsymbol{y} - 2\beta^T\boldsymbol{X}^T\boldsymbol{y} + \beta^T\boldsymbol{X}^T\boldsymbol{X}\beta\end{aligned}$$

We optimise the RSS by taking the $p + 1$ partial derivatives and setting them to zero.

Vector differentiation rules

Let $\boldsymbol{x}$ and $\boldsymbol{a}$ be vectors of equal dimension and $\boldsymbol{A}$ a matrix with column dimension the same as number of rows in $\boldsymbol{x}$. Then:

$$\frac{d(\boldsymbol{x}^T\boldsymbol{a})}{d\boldsymbol{x}} = \boldsymbol{a}$$

$$\frac{d(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x})}{d\boldsymbol{x}} = (\boldsymbol{A} + \boldsymbol{A}^T)\boldsymbol{x}$$

Partial derivatives

$$RSS(\beta) = \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T\boldsymbol{X}\beta + \beta^T\boldsymbol{X}^T\boldsymbol{X}\beta$$

The vector of partial derivatives is

$$\frac{d(\text{RSS}(\beta))}{d\beta} = \frac{d(\boldsymbol{y}^T\boldsymbol{y})}{d\beta} - \frac{d(2\beta^T\boldsymbol{X}^T\boldsymbol{y})}{d\beta} + \frac{d(\beta^T\boldsymbol{X}^T\boldsymbol{X}\beta)}{d\beta} =$$

$$= \mathbf{0} - 2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\beta$$

OLS estimates

The first order condition is:

$$\frac{d(\text{RSS}(\boldsymbol{\beta}))}{d\boldsymbol{\beta}} = -2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} = \mathbf{0}$$

The least squares estimate $\widehat{\boldsymbol{\beta}}$ therefore satisfies

$$\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{y}$$

If $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ is invertible, left multiplication with $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ gives the unique solution

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}.$$

Fitted values

The fitted values based on the training inputs are

$$\widehat{y}_i = \widehat{\beta}_0 + \sum_{j=1}^{p} \widehat{\beta}_j x_{ij}$$

The vector of fitted values for the entire sample is:

$$\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}.$$

We refer to $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ as the **hat matrix**.

Residuals

The regression **residuals** are:

$$\begin{aligned} e_i &= y_i - \widehat{y}_i \\ &= y_i - \widehat{\beta}_0 - \sum_{j=1}^{p} \widehat{\beta}_j x_{ij} \end{aligned}$$

The vector of residuals is:

$$\begin{aligned} \boldsymbol{e} &= \boldsymbol{y} - \widehat{\boldsymbol{y}} \\ &= \boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \\ &= \left(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\right)\boldsymbol{y}. \end{aligned}$$

Measuring fit

We can show that

$$\text{TSS} = \text{RegSS} + \text{RSS}$$

$$\sum (y_i - \overline{y})^2 = \sum (\widehat{y}_i - \overline{y})^2 + \sum e_i^2$$

- TSS: total sum of squares.
- RegSS: regression sum of squares.
- RSS: residual sum of squares.

Measuring fit

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

Interpretation:

- The R^2 measures the proportion of the variation in the response data that is accounted for by the estimated linear regression model.
- The R^2 can only increase when you add another variable to the model.
- The R^2 is an useful part of the regression toolbox, but it does not measure the predictive accuracy of the estimated regression, or more generally how good the model is.

Prediction

Let $\widehat{\beta}$ be the OLS coefficients obtained from the training sample.

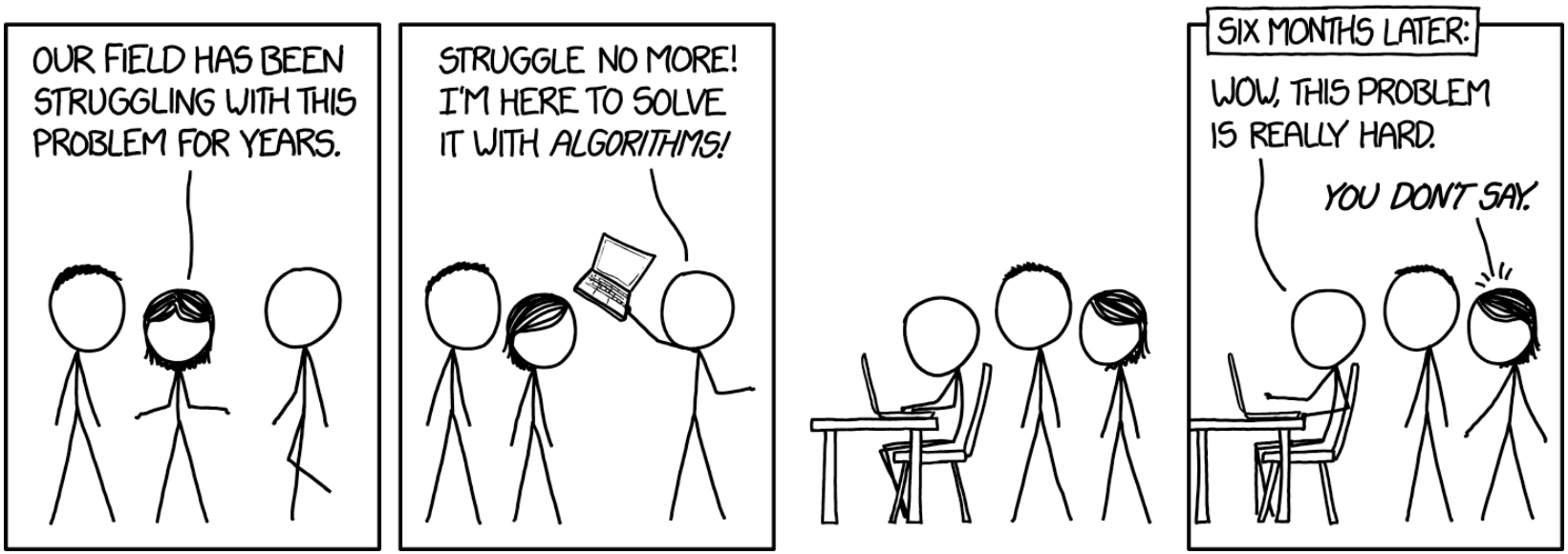
$$\widehat{y}_0 = \widehat{\beta}_0 + \sum_{j=1}^{p} \widehat{\beta}_j x_{0j}$$

These predictions are weighted sums over the training data.

The MLR model

Models and algorithms

So far, we have talked about the least squares algorithm and even arrived at predictions without reference to a model. The current practice of data science places large emphasis on **algorithmic thinking** towards problem solving.



https://xkcd.com/1831/

Statistical models

A **statistical model** is a description of a data generating process based on a set of mathematical assumptions about the population and the sampling process.

A **regression model** is a description of the relationship between a response variable Y and predictors $X_1, \ldots, X_p$. More formally, it is a model of the form $p(y|\boldsymbol{x}; \theta)$.

Formulating statistical models and making assumptions allow us to say more about a problem.

The Multiple Linear Regression (MLR) model

1. Linearity: if $X = \boldsymbol{x}$, then

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon$$

   for some population parameters $\beta_0, \beta_1, \ldots, \beta_p$ and a random error ε.

2. The conditional mean of ε given X is zero: $E(\varepsilon|X) = 0$.

3. Constant error variance: $\text{Var}(\varepsilon|X) = \sigma^2$.

4. Independence: all the error pairs ε_i and ε_j $(i \neq j)$ are independent.

5. The distribution of $X_1, \ldots, X_p$ is arbitrary.

6. There is no perfect multicollinearity.

Checking the assumptions

It is fundamental to check the assumptions with data. We do this with **residual diagnostics**. The following plots are often useful:

- Fitted values against residuals.
- Predictors against residuals.
- Fitted values against squared or absolute residuals.
- Predictors against squared or absolute residuals.
- Residual distribution.
- If the observations are ordered: residuals against coordinates (time and/or space).

Statistical properties

Sampling distribution of an estimator

In classical statistics, the population parameter β is fixed and the data is a random sample from the population. We estimate β by applying an **estimator** $\widehat{\beta}(\mathcal{D})$ to data (in our case the OLS algorithm).

We study the uncertainty of an estimate by computing the **sampling distribution** of the estimator.

Sampling distribution of an estimator

Imagine that we draw many different datasets $\mathcal{D}^{(s)}$ $(s = 1, \ldots, S)$ from the true model $p(\boldsymbol{y}|\boldsymbol{X}; \boldsymbol{\beta})$. Each dataset has size N.

For each of these datasets, we apply the estimator $\widehat{\boldsymbol{\beta}}(\cdot)$ and obtain a set of estimates $\{\widehat{\boldsymbol{\beta}}(\mathcal{D}^{(s)})\}$. The sampling distribution is the induced distribution on $\widehat{\boldsymbol{\beta}}(\cdot)$ as $S \to \infty$.

This concept is not necessarily intuitive since it refers to hypothetical datasets rather than data that we do have.

Sampling distribution of an estimator

Establishing the sampling distribution allows us to answer questions such as:

- Is there are relationship between the response and the predictors?
- Are all the predictors related to the response, or only a subset?
- How accurate are our coefficient estimates?
- How accurate are our predictions?

Mean and variance

Based on the MLR model assumptions, the mean and variance of the OLS estimator are

$$E(\widehat{\boldsymbol{\beta}}|\boldsymbol{X}) = \boldsymbol{\beta}$$

$$\text{Var}\left(\widehat{\boldsymbol{\beta}}|\boldsymbol{X}\right) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

For individual coefficients,

$$\text{Var}\left(\widehat{\beta}_j\right) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{jj},$$

where $(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{jj}$ is the jth diagonal element of $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$.

Variance

We can obtain the more interpretable expression

$$\text{Var}(\widehat{\beta}_j) = \frac{\sigma^2}{\left(1 - \text{R}_j^2\right)\left(\sum_{i=1}^{N}(x_{ji} - \overline{x}_j)^2\right)} = \frac{\sigma^2}{\left(1 - \text{R}_j^2\right)(n-1)s_{x_j}^2}$$

R_j^2: the r-squared of a regression of predictor j on all other predictors.

x_{ij}: observed value of predictor j for observation i.

$\overline{x}_j$: sample average of predictor j.

$s_{x_j}^2$: sample variance of predictor j.

Variance

$$\text{Var}(\widehat{\beta}_j) = \frac{\sigma^2}{\left(1 - \text{R}_j^2\right)\left(\sum_{i=1}^{N}(x_{ij} - \overline{x}_j)^2\right)}$$

Interpretation:

- The higher the correlation of predictor j with other predictors, the higher the variance of $\widehat{\beta}_j$.
- The variance of $\widehat{\beta}_j$ is proportional to the variance of errors σ^2.
- The variance decreases with the sample variance of predictor j.
- The variance decreases with N.

Standard errors

We estimate the variance of the errors σ^2 as

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^{N} e_i^2}{N - p - 1}$$

That leads to the standard errors for each individual estimate

$$\mathsf{SE}(\widehat{\beta}_j) = \widehat{\sigma}\sqrt{(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{jj}},$$

Desirable properties of estimators

- An estimator is **consistent** when it recovers the true parameter that generated the data when sample size goes to infinity.

- An estimator $\widehat{\theta}(\mathcal{D})$ is **unbiased** when $E(\widehat{\theta}) = \theta$, where θ is the true parameter.

- Minimum variance (efficiency).

We can show that under the MLR assumptions, the OLS estimator is unbiased and consistent. Furthermore, it is has the lowest variance among all linear unbiased estimators (Gauss-Markov Theorem).

Predictions

Under the MLR assumptions, the OLS method leads to unbiased predictions. Furthermore, the variance of the prediction errors is

$$E\left[\left(Y_0 - \boldsymbol{x}_0^T\widehat{\beta}\right)^2\right] = \sigma^2 + \text{Var}\left(\boldsymbol{x}_0^T\widehat{\beta}\right)$$
$$= \sigma^2(1 + \boldsymbol{x}_0^T\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{x}_0)$$

Interpretation: $\boldsymbol{x}_0^T\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{x}_0$ is a Mahalanobis distance. The predictions are less accurate when $\boldsymbol{x}_0$ is away from the center of the predictor data.

Optimal linear prediction

- Our discussion has been based on strong assumptions about the data. In particular, the assumption that $E(\varepsilon|X) = 0$ in particular is unlikely to hold for observational data. Moreover, linearity is likely to be at best an approximation.

- We can drop most assumptions to show that the OLS leads to optimal linear predictions under the squared error loss when the size of the training data goes to infinity.

The Gaussian MLR model

The Gaussian MLR model

- Our analysis so far did not make any **distributional assumptions** about the regression errors. Even though we assumed conditions such as $E(\varepsilon|X) = 0$ and $\text{Var}(\varepsilon|X) = \sigma^2$, we left the probability distribution of ε unspecified.

- We managed to learn quite a lot from these minimal assumptions. For example, the assumptions for the conditional mean and variance of the errors naturally leads us to the mean and variance of the OLS estimator.

- But we may want to learn more. For example, what is the full sampling distribution of the OLS estimator? Knowing this distribution is necessary for making probability statements about the uncertainty in this estimator.

The Gaussian MLR model

We now add the assumption that $\varepsilon \sim N(0, \sigma^2)$, leading to the model equation

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j, \qquad \varepsilon_i \sim N(0, \sigma^2).$$

A key feature of the Gaussian MLR model is that it gives us the full form of the conditional distribution of Y:

$$Y|X = \boldsymbol{x} \sim N\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j,\, \sigma^2\right)$$

Maximum Likelihood Estimation

- **Maximum likelihood** (ML) estimation is available when we specify a full probabilistic model for the population. We now introduce this concept for the specific case of the Gaussian MLR model.

- Intuitively, ML estimation chooses the values of the parameters that maximise the likelihood of the observed data under the model (for discrete data the likelihood is the probability of the data, but our response Y is continuous).

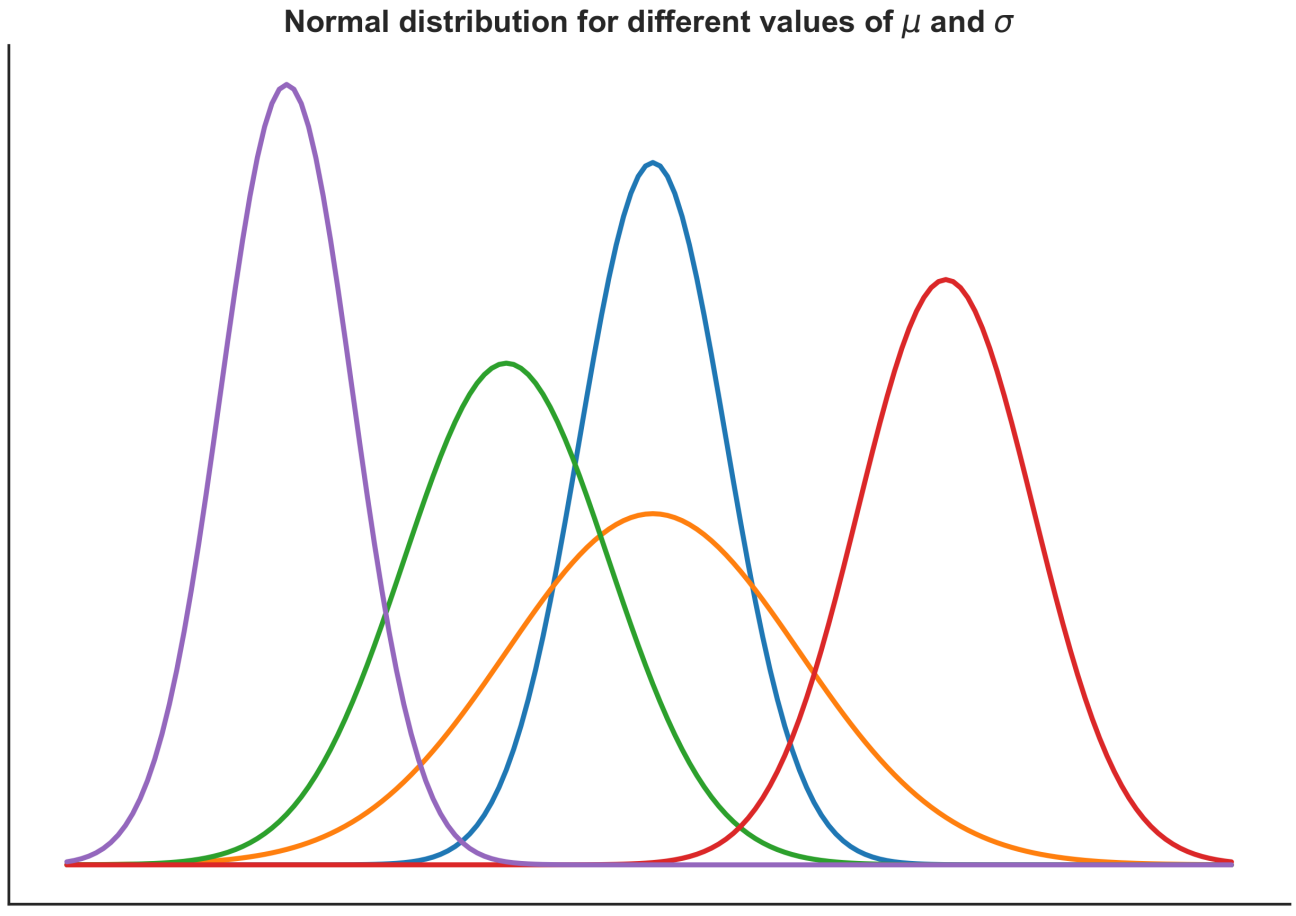
- ML is one of the most highly used estimation techniques in statistics.

Maximum Likelihood Estimation

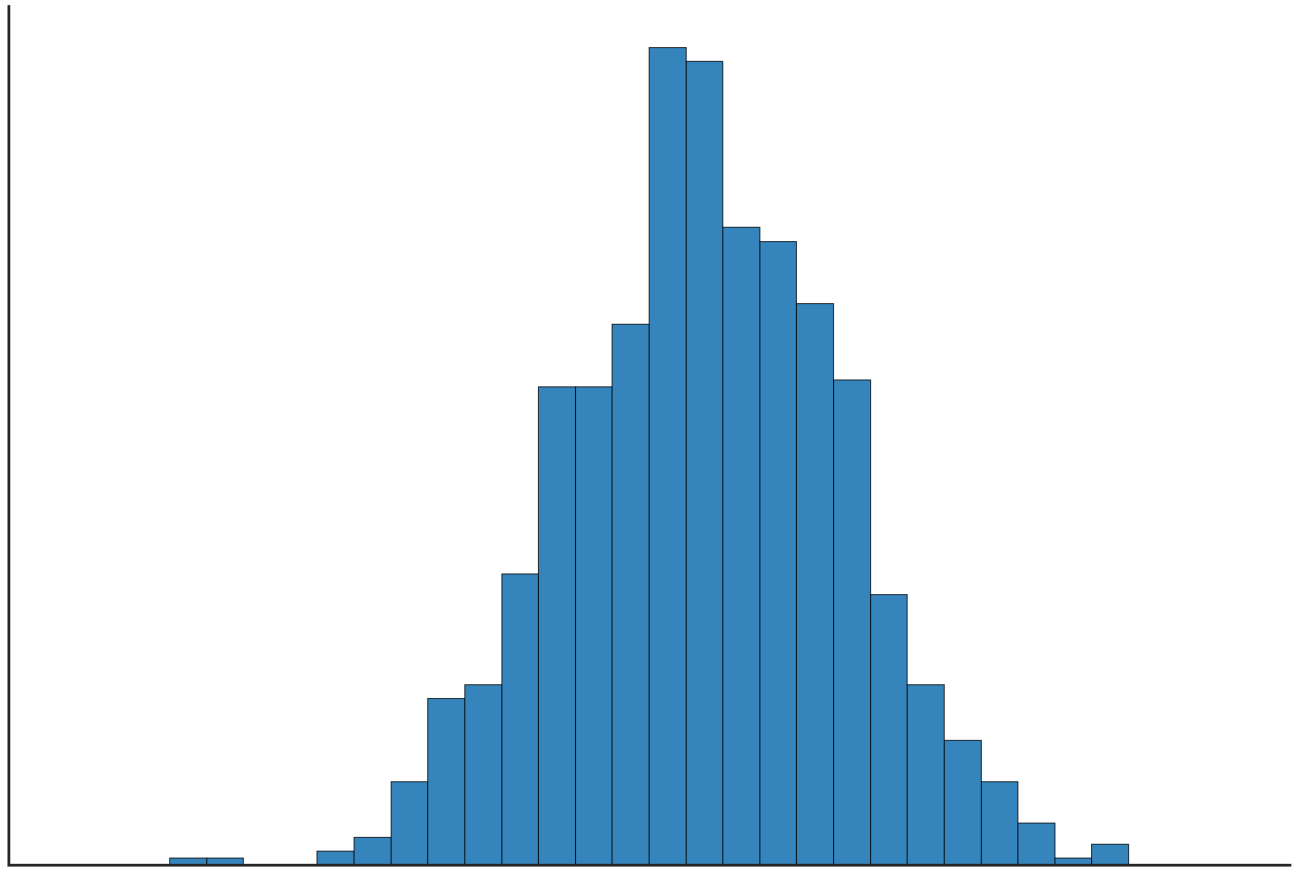
Recall the formula for the normal probability density function (PDF) from basic statistics:

$$p(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

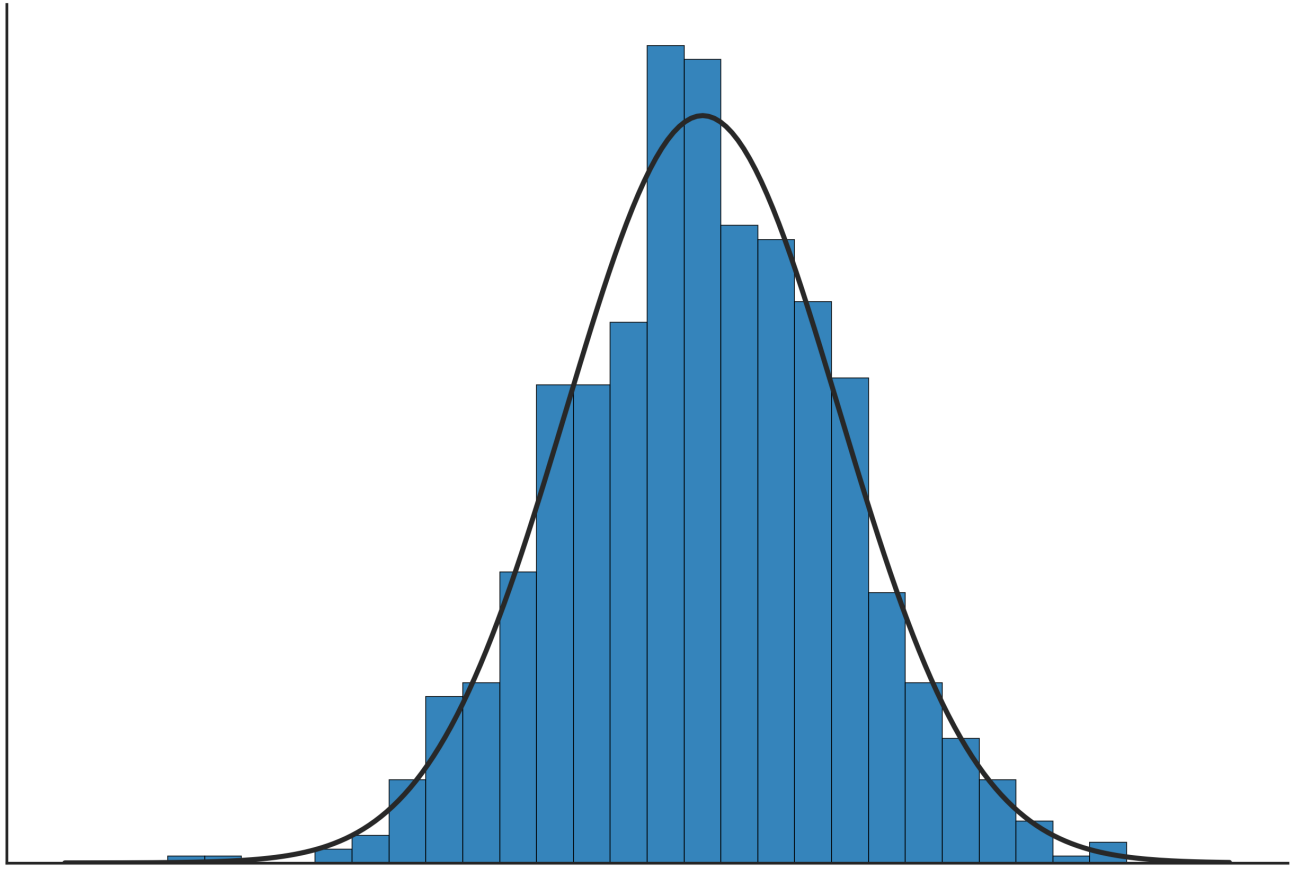
Maximum Likelihood Estimation



Maximum Likelihood Estimation



Maximum Likelihood Estimation



Maximum Likelihood Estimation

Since

$$Y_i|X_i = \boldsymbol{x}_i \sim N\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}, \sigma^2\right),$$

the density for an observed value y_i is

$$p(y_i|x_i; \boldsymbol{\beta}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp^{-\frac{(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2}{2\sigma^2}}$$

Maximum Likelihood Estimation

The **likelihood function** is the joint PDF of the data evaluated at the sample values. In our Gaussian SLR model, Assumption 4 (independence) implies that we can multiply the PDFs for each observation:

$$p(\boldsymbol{y}; \boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp^{-\frac{\left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}\right)^2}{2\sigma^2}}$$

Maximum likelihood estimation

The **log-likelihood** is the log-density of the observed sample,

$$\mathcal{L}(\boldsymbol{\beta}, \sigma^2) = \log \prod_{i=1}^{N} p(y_i; \boldsymbol{\beta}, \sigma^2) = \sum_{i=1}^{N} \log p(y_i; \boldsymbol{\beta}, \sigma^2)$$

$$= -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2$$

Maximum likelihood estimation

We maximise the log-likelihood as a function of the parameters

$$\max_{\boldsymbol{\beta},\sigma^2} \; \mathcal{L}(\boldsymbol{\beta},\sigma^2),$$

where

$$\mathcal{L}(\boldsymbol{\beta},\sigma^2) = -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2$$

Note that the last term corresponds RSS($\boldsymbol{\beta}$) times a negative multiplier. Therefore, ML estimator is equivalent to the OLS estimator for this model.

Discussion

We ended up back at OLS, so what is the take-away?

- We added a new estimation principle to to our toolbox and started by understanding it in this sample case.
- ML estimation is broadly applicable and we will used extensively for supervised learning.
- We need the concept of a log-likelihood for certain model selection methods.

Discussion

When discussing statistical decision theory, we defined the optimal prediction rule

$$\delta(\boldsymbol{x}) = \underset{m}{\text{argmin}}\ E(L(Y, m)|X = \boldsymbol{x}),$$

A probabilistic model estimated by ML will allow us to directly approximate the optimal prediction for any loss, since it estimates the full conditional distribution $P(Y|X = \boldsymbol{x})$.

This is therefore a different approach compared estimation by minimising the empirical risk for the training sample. That leads to the natural question: which one is better?

Sampling distribution

Under the Gaussian MLR model, we can obtain an exact sampling distribution for the OLS estimator,

$$\widehat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right)$$

When estimating σ^2, we have

$$\frac{\widehat{\beta}_j - \beta_j}{\mathsf{SE}(\widehat{\beta}_j)} \sim t_{N-p-1},$$

where t_{N-p-1} denotes the Student's t distribution with $N-p-1$ degrees of freedom. We can then rely on this distribution for hypothesis testing.

Confidence interval

Recall the basic structure:

point estimate $\pm$ **critical value** $\times$ **standard error**

$$\frac{\widehat{\beta}_j - \beta_j}{\mathsf{SE}(\widehat{\beta}_j)} \sim t_{N-p-1}$$

The approximate $100 \times (1-\alpha)\%$ confidence interval for β_j is

$$\widehat{\beta}_j \pm t_{N-p-1,\alpha/2} \times \mathsf{SE}(\widehat{\beta}_j)$$

Hypothesis testing for multiple coefficients

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$$
$$H_1 : \text{at least one } \beta_j \neq 0$$

$$F_{\text{stat}} = \frac{(\text{TSS} - \text{RSS})/p}{\text{RSS}/(N-p-1)} = \frac{\text{R}^2}{1-\text{R}^2} \times \frac{(N-p-1)}{p},$$

$$F_{\text{stat}} \sim F_{p,\, N-p-1}$$

Hypothesis testing for multiple coefficients

Suppose that we want to test whether q out of the p coefficients are zero, or more generally q linear restrictions.

$$\mathsf{F}_{\text{stat}} = \frac{(\mathsf{RSS}_0 - \mathsf{RSS})/q}{\mathsf{RSS}/(N-p-1)},$$

where RSS_0 and RSS are the residual of squares of the null and full models respectively.

Equivalently,

$$\mathsf{F}_{\text{stat}} = \frac{(\mathsf{R}^2 - \mathsf{R}_0^2)/q}{\left(1-\mathsf{R}^2\right)/(N-p-1)}$$

Under H_0 (the restricted model), $\mathsf{F}_{\text{stat}} \sim F_{q,\,N-p-1}$.

Interpreting a linear regression model

Advertisement data

We now estimate the linear regression model

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

To interpret the results, we need to note the following:

- The observational units in the data are markets.
- The response variable (sales) is in thousands of units.
- The predictors are in thousands of dollars.

What is the population of interest? (You always need to be able to answer this question)

Advertisement data

OLS Regression Results

Dep. Variable:	Sales	R-squared:	0.897
Model:	OLS	Adj. R-squared:	0.896
Method:	Least Squares	F-statistic:	570.3
Date:		Prob (F-statistic):	1.58e-96
Time:		Log-Likelihood:	-386.18
No. Observations:	200	AIC:	780.4
Df Residuals:	196	BIC:	793.6
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.9389	0.312	9.422	0.000	2.324	3.554
TV	0.0458	0.001	32.809	0.000	0.043	0.049
Radio	0.1885	0.009	21.893	0.000	0.172	0.206
Newspaper	-0.0010	0.006	-0.177	0.860	-0.013	0.011

Omnibus:	60.414	Durbin-Watson:	2.084
Prob(Omnibus):	0.000	Jarque-Bera (JB):	151.241
Skew:	-1.327	Prob(JB):	1.44e-33
Kurtosis:	6.332	Cond. No.	454.

Interpreting coefficients

$$\widehat{\text{sales}} = \underset{(0.312)}{2.94} + \underset{(0.001)}{0.046} \times \text{TV} + \underset{(0.009)}{0.189} \times \text{radio} - \underset{(0.006)}{0.0010} \times \text{newspaper}$$

Interpretation (TV):

If we select two markets from the population, where the radio and newspaper budgets are the same, but the TV budget differs by 100 dollars, we would expect 4.6 more units sold in the market with higher TV budget.

Interpreting coefficients

Mathematically:

$$\beta_j = E(Y|X_j = x_j + 1, X_{\neq j} = \boldsymbol{x}_{\neq j}) - E(Y|X_j = x_j, X_{\neq j} = \boldsymbol{x}_{\neq j})$$

For example, with $p = 2$ and focusing on the first predictor:

$$\begin{aligned}
&E(Y|X_1 = x_1 + 1, X_2 = x_2) - E(Y|X_1 = x_1, X_2 = x_2) \\
&= E[\beta_0 + \beta_1(x_1 + 1) + \beta_2 x_2 + \varepsilon] - E[\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon] \\
&= [\beta_0 + \beta_1(x_1 + 1) + \beta_2 x_2] - [\beta_0 + \beta_1 x_1 + \beta_2 x_2] \\
&= \beta_1
\end{aligned}$$

Interpreting coefficients

$$\widehat{\text{sales}} = \underset{(0.312)}{2.94} + \underset{(0.001)}{0.046} \times \text{TV} + \underset{(0.009)}{0.189} \times \text{radio} - \underset{(0.006)}{0.0010} \times \text{newspaper}$$

Wrong interpretation (unless supported by the study design):

If we increase the TV by 100 dollars, we expect sales to increase by 4.6 units (holding radio and newspaper fixed).

Omitted variables

The assumption that $E(\varepsilon|X = \boldsymbol{x}) = 0$ is generally not satisfied for observational data. This is due to the presence of **omitted variables**, variables that are correlated with both the predictor and the response. The estimated regression coefficients are then subject to **omitted variable bias**.

Here is an example: if we regress wealth on the number of luxury cars owned, the slope is positive (luxury cars predict wealth). However, we can imagine that buying more luxury cars will not make you richer.

Example: education and wages

OLS Regression Results

Dep. Variable:	Hourly wage	R-squared:	0.162
Model:	OLS	Adj. R-squared:	0.162
Method:	Least Squares	F-statistic:	1729.
Date:		Prob (F-statistic):	0.00
Time:		Log-Likelihood:	-57425.
No. Observations:	17919	AIC:	1.149e+05
Df Residuals:	17916	BIC:	1.149e+05
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-7.5017	0.337	-22.278	0.000	-8.162	-6.842
Education	1.1937	0.024	50.255	0.000	1.147	1.240
Experience	0.4511	0.011	40.772	0.000	0.429	0.473

Omnibus:	10774.032	Durbin-Watson:	0.744
Prob(Omnibus):	0.000	Jarque-Bera (JB):	237446.384
Skew:	2.484	Prob(JB):	0.00
Kurtosis:	20.128	Cond. No.	117.

Causal analysis

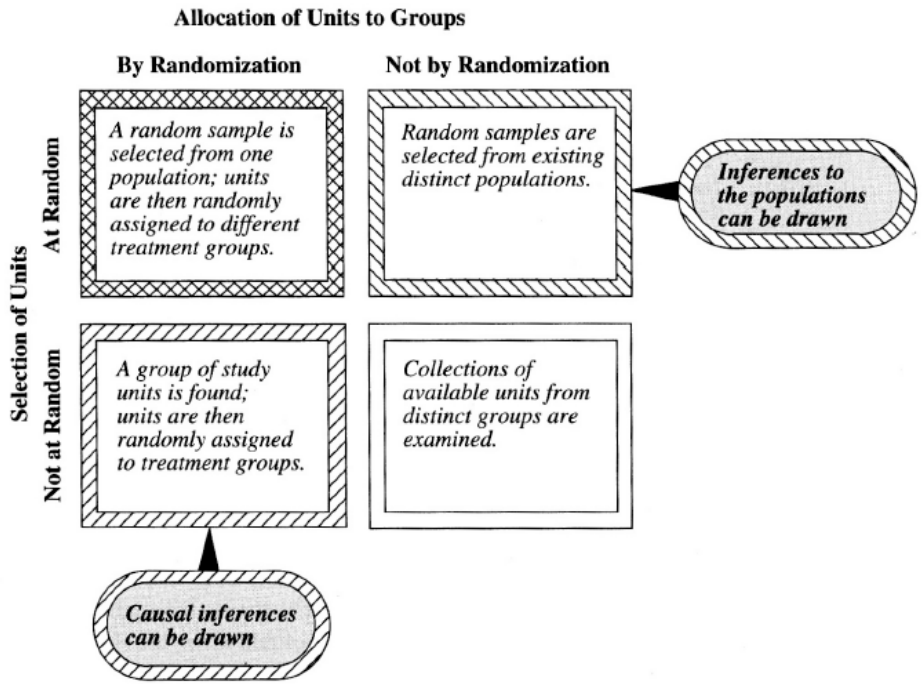
Causal analysis means to estimate a model of the type $E(Y|\text{ do } X = \boldsymbol{x})$. This is an explicit intervention: "if we do $X = \boldsymbol{x}$, then we predict $E(Y|X = \boldsymbol{x})$".

This is different from predictive modelling: "if we observe $X = \boldsymbol{x}$, then we predict $E(Y|X = \boldsymbol{x})$".

Causal analysis requires an appropriate **study design** (such as A/B testing).

Study designs

Display 1.5 Statistical inferences permitted by study designs



Ramsey and Shafter (2002).

Other interpretation mistakes

- Saying that a predictor is important because the coefficient is statistically significant.

- Saying that a predictor is not important because the coefficient is not statistically different from zero.

Study note

- For our purposes, the textbook is not sufficiently rigorous regarding the interpretation of linear regression coefficients.
- While our interpretation is less simple than the one provided by most textbooks, it is the correct one for observational data that is prevalent in business.

Review questions

- How do we obtain the OLS estimates? Go through the full process. (Optional for advanced students.)
- What is a sampling distribution?
- What is maximum likelihood estimation? What type of model is it applicable to?
- We formulated several questions about the advertisement data. Answer some of these questions based on the output in page 56.
- What is the correct interpretation of a linear regression model coefficient with observational data?
- What is the difference between predictive and causal analysis?