Lecture notes for MA317-6-AU

Dr. Danilo Petti
Department of Mathematical Sciences
University of Essex

2023-2024

Contents

Declaration

These lecture notes do not represent original content; their content is extracted from the references cited in the appropriate section. Furthermore, this material is intended to provide a comprehensive overview of the topics covered in the course and are the result of research and analysis. While they may not cover every detail of the course material, they are designed to give students a solid understanding of the key concepts and principles. It should be noted that the lecture notes are not a substitute for attending lectures and actively participating in class discussions, but rather a complement to them. The notes serve as a tool for students to review and reinforce their understanding of the material presented in class, and the exercises assigned on a weekly basis are part of the lecture material and should be attempted by students to fully grasp the concepts covered. Students are encouraged to consult the references provided and conduct further research to deepen their knowledge on the topics covered in the course. The instructor and teaching assistants are available to provide guidance and support throughout the course. Finally, it is important to adhere to academic integrity principles and properly cite any sources used in research and assignments.

Module assessment

lab test 30% and final exam 70%

References

- Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to linear regression analysis* (6th ed.). Wiley.
  - **Note for students:** Sections 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 13.
- Verbeek, M. (2017). *A guide to modern econometrics* (5th ed.). Wiley.
  - **Note for students:** Sections 1, 2, 3, 6, Appendix A & B.
- Faraway, J. J. (2015). *Linear Models with R* (2nd ed.). CRC Press.
  - **Note for students:** Covers the material discussed in the labs.
- Faraway, J. J. (2016). *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models* (2nd ed.). CRC Press.
- Data used in the Labs are mainly extracted from: Faraway J (2022). _faraway: Functions and Datasets for Books by Julian Faraway_. R package version 1.0.8, <https://CRAN.R-project.org/package=faraway>.

Syllabus

Simple linear regression

- Explain what is meant by response and explanatory variables.
- Explain what is meant by a variable, a factor taking categorical values and an interaction term.
- Use simple linear regression to describe the stages of conducting a data analysis to solve real-world problems in a scientific manner and describe tools suitable for each stage.
- Use simple linear regression to describe sources of data and explain the characteristics of different sources, including extremely large data sets.
- Explain the meaning and value of reproducible research and describe the elements required to ensure a data analysis is reproducible.
- Maximum likelihood estimation.
- Link between maximum likelihood and least Squares. OLS for linear regression.
- Use a fitted linear relationship for prediction. Define the linear predictor.
- Confidence intervals for parameters and prediction intervals for future observations.
- ANOVA

Multiple linear regression

- Multiple regression. Subdividing the regression sum of squares. Lack of fit and pure error.
- Regression diagnostics, violation of the assumptions, leverage points, outliers and collinearity and methods to deal with them.
- Apply statistical tests to determine the acceptability of a fitted model: Testing one linear restriction, joint test of significance of regression, general case.
- Principal Component Regression
- Principal Component Analysis
- Model selection via stepwise methods. Cp, AIC, BIC
- Ridge regression
- Lasso Regression

Logistic regression

- Define the deviance and scaled deviance and state how the parameters of a logistic model may be estimated.
- Define the Pearson and deviance residuals and describe how they may be used.
- Use R to implement the methods above on a data set and interpret the output.

Chapter 1

The Method of Maximum Likelihood

Suppose that random variables $X_1, \ldots, X_n$ have a joint density or frequency function defined as $f(x_1, \ldots, x_n|\theta)$. Given observed sample $(x_1, \ldots, x_n)$, the likelihood function of θ is defined as

$$L(\theta|\mathbf{x}) = f(x_1, \ldots, x_n|\theta), \;\; \theta \in \Theta, \tag{1.1}$$

where Θ denotes the parametric space. We can note that the function just presented is not the joint density, this is because it is no longer a function of the random variables X, this because we observed a sample and we are evaluating (**??**) according to $(x_1, \ldots, x_n)$. The maximum likelihood estimate (MLE) of θ is that value of θ that maximizes the likelihood function, the value that makes the observed data most likely to be observed.

If the X_i are assumed to be i.i.d., then the joint density can be written as the product of the marginal densities, and the likelihood function is

$$L(\theta|\mathbf{x}) = \prod_{i=1}^{n} f_{X_i}(x_i|\theta).$$

Rather than maximizing the likelihood, it is usually easier (and equivalent) to maximize its natural logarithm. The log-likelihood function is

$$\ell(\theta) = \sum_{i=1}^{n} \log[f_{X_i}(x_i|\theta)].$$

Let us find the MLE for the Gaussian distribution

Example 1. *(Gaussian Distribution)*

If $X_1, \ldots, X_n$ are i.i.d. $\mathcal{N}(\mu, \sigma^2)$, their joint density is the product of their marginal densities:

$$f(x_1, \ldots, x_n|\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(x_i - \mu)^2}{2\sigma^2} \right\},$$

the log-likelihood is

$$\ell(\mu, \sigma^2) = -n \log \sigma - \frac{n}{2} \log 2\pi - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2,$$

the partials derivatives respect to μ and σ are

$$\begin{aligned} \frac{\partial \ell(\mu, \sigma^2)}{\partial \mu} &= \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu), \\ \frac{\partial \ell(\mu, \sigma^2)}{\partial \sigma} &= -\frac{n}{\sigma} + \sigma^{-3} \sum_{i=1}^{n} (x_i - \mu)^2, \end{aligned}$$

setting the first and second partial equal to zero and solving for μ and σ we obtain the MLEs

$$\begin{aligned} \hat{\mu} &= \bar{X}, \\ \hat{\sigma} &= \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{X})^2}, \end{aligned}$$

where $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} x_i$

Example 2. *If X follows a Poisson distribution with paramter λ then*

$$\mathbb{P}(X = x) = \frac{\lambda^x e^{-\lambda}}{x!},$$

If $X_1, \ldots, X_n$ are i.i.d. and Poisson, their joint distribution function is the product of the marginal distributions. The log-likelihood is

$$\begin{aligned} \ell(\lambda) &= \sum_{i=1}^{n} (x_i \log \lambda - \lambda - \log x!), \\ &= \log \lambda \sum_{i=1}^{n} x_i - n\lambda - \sum_{i=1}^{n} \log x_i!. \end{aligned}$$

Setting the first derivative equal to zero and solving for λ, we end up with

$$\left. \frac{\partial \ell(\lambda)}{\partial \lambda} \right|_{\lambda = \hat{\lambda}} = \hat{\lambda}^{-1} \sum_{i=1}^{n} x_i - n = 0,$$

the MLE is

$$\hat{\lambda} = \bar{X}.$$

A very useful quantity in the context of the maximum likelihood estimation is the Fisher information Marix with the jk^{th} $(1 \leq, k \leq d)$ entry is defined as

$$i_{jk}(\theta) = \mathbb{E}\left[\frac{\partial^2}{\partial\theta_j \partial\theta_k}\ell(\theta)\right].$$

This can be thought of as a measure of how hard it is to estimate θ when it is the true parameter value. The Cramer-Rao Lower bound states that if $\hat{\theta}$ is an unbiased estimator of θ, the under regularity conditions

$$Var(\hat{\theta}) - i^{-1}(\theta),$$

is a positive definite quantity. In other words we are saying the MLE achieves the lowest possible variance.

The main asymptotic properties of maximum likelihood estimators are

- Under regularity conditions, the MLE ($\hat{\theta}$) is a strongly consistent estimator of θ

$$\mathbb{P}\left(\lim_{n\to\infty} \hat{\theta} = \theta\right) = 1,$$

  it follows that the MLE is asymptotically unbiased, since

$$\mathbb{E}(\hat{\theta} - \theta) \to 0 \text{ as } n \to \infty.$$

- The maximum likelihood estimator is asymptotically efficient. That is

$$\lim_{n\to\infty} \text{Var}(\hat{\theta}) = \mathcal{I}(\theta)^{-1}$$

- The MLE is asymptotically normally distributed and efficient

$$\sqrt{n}(\hat{\theta} - \theta) \sim \mathcal{N}(0, \mathcal{I}(\theta)^{-1}),$$

  where $\mathcal{I}(\theta)$ is the fisher information matrix

with regularity conditions we mean

- The likelihood function is continuous in θ;
- The density/mass function is such that for all x where $f(x;\theta) > 0$

$$\frac{\partial}{\partial\theta}\log f(x,\theta),$$

exists and is finite;

- The order of differentiation with respect to θ and integration over the sample space may be reversed. In practice, we have this regularity condition satisfied where the support of $f(x, \theta)$ does not depend on θ.

Theorem 1. *Let $\hat{\theta}$ the MLE for the paratemeter θ and $\psi(\cdot)$ a one-to-one function, then $\psi(\hat{\theta})$ is the MLE for $\psi(\theta)$.*

Exercises

1. Let $x_1, x_2, \ldots, x_n$ be a simple random sample drawn from a random variable X with density function:

$$f(x;\lambda) \propto \frac{x}{\frac{2}{3}\lambda^2} e^{-\frac{x}{\lambda}}$$

for $x > 0$ and $\lambda > 0$.

The candidate should:

(a) Define the maximum likelihood function and the log-likelihood function;

(b) Calculate the maximum likelihood estimate $\hat{\lambda}$ for λ;

2. Suppose we have a single observation $X = x$ from an exponential distribution with parameter λ and pdf

$$f(x|\lambda) = \lambda e^{-\lambda x}, \quad x > 0.$$

Find the maximum likelihood estimate (MLE) $\hat{\lambda}$ of λ. Now reparametrize using $\theta = \frac{1}{\lambda}$ and find the MLE $\hat{\theta}$ of θ. Comment on the relationship between $\hat{\lambda}$ and $\hat{\theta}$. Show that $\hat{\theta}$ is unbiased for θ and find its variance. Is $\hat{\lambda}$ unbiased for λ?

Useful integrals: for $n \geq 1$

$$\int_0^\infty y^{n-1} e^{-y} dy = \Gamma(n) = (n-1)! \tag{1}$$

and

$$\int_0^\infty y^{-1} e^{-y} dy = \infty. \tag{2}$$

3. With θ as the parameter so that

$$f(x|\theta) = \frac{1}{\theta} e^{-x/\theta}, \quad x > 0,$$

suppose we now have n independent observations $x_1, x_2, \ldots, x_n$. Find the MLE of θ, show it is unbiased and find its variance.

4. The truncated exponential distribution has pdf

$$f(x|\theta, \alpha) = \frac{1}{\theta} e^{-(x-\alpha)/\theta}, \quad x > \alpha.$$

Sketch the pdf and check that it satisfies the two conditions for being a pdf. For known α and n independent observations $x_1, x_2, \ldots, x_n$, what is the MLE of θ? Now suppose both α and θ need to be estimated. Find their MLEs.

Hint: You need to be careful with α; think before you differentiate.

Chapter 2

Simple Linear Regression

The most operative part of statistics is about the model building. The main goal of an analyst is to *describe*,*understand*,*make prediction*, *simulate* and *control* a phenomenon. For these purposes, it is crucial to understand the logical and formal structure of a regression model. In which the final goal is to define an explicit relationship between what is meant to be explained Y and its cause X.

A statistical model is by definition a simplified representation of reality. Which derives from sample observations and logical deductions. The construction of a statistical model is characterized by three main phases: **specification, estimation** and **validation phase**.

i) **Model specification**, in this phase our aim is to to specify a functional form between variables of interest, in the following way

$$Y = f(X_1, \ldots, X_p),$$

where Y is the dependent variable (outcome variable, response variable) and $X_1, \ldots, X_p$ are the independent variables (predictor variables, explanatory variables). We are going to use X's to explain Y through $f(\cdot)$.

The relationship $Y = f(X_1, \ldots, X_p)$ can usually be derived from two main factors i) prior knowledge of the phenomenon ii) experimental results. In the specification phase, the variables of interest and their role must be identified, because this helps to formulate the functional link $f(\cdot)$ correctly.

Any statistical model that we are going to encounter in this module represents an approximation of the reality. As there is no scientific field in which is legitimate to hypothesize a deterministic relationship among random variables, we need to take another component in our specification. A component that helps taking into account the uncertainty. Therefore

$$Y = f(X_1, \ldots, X_p) + \epsilon,$$

where ϵ is a random variable (error term) that represents our *real or supposed ignorance* about

the relationship among Y and $X_1, \ldots, X_p$. In real word applications, the specification of the functional form $f(\cdot)$ comes immediately from the nature of the problem or the underlying theory (Examples: marco and micro-economics, medicine, biology, chemistry).

- If Y is the weight and X is the height, a reasonable functional form is: $Y = \beta X + \epsilon$ where $\beta \in \mathbb{R}$;
- If Y is the weight of a tile and X_1 and X_2 represent the length and width, knowing that the weight of a tile is proportional to its area a functional form is: $Y = \beta X_1 X_2 + \epsilon$;
- The micro-economic theory teaches us that the output Y depends from the capital K and labour force L through the Cobb–Douglas production function: $Y = \beta_0 K^{\beta_1} L^{\beta_2} + \epsilon$ where $\beta_0, \beta_1, \beta_2 \in \mathbb{R}$.

In each of the proposed specifications, one or more parameters β need(s) to be estimated in order to use these models. Even if these parameters assume specific meanings (e.g., constants of proportionality, basis weight, elasticity of demand, rate of change), this is not necessary for the purpose of the specification phase.

A statistical model can be classified using different criteria, according to the data employed, the estimation method, the nature of the random variables, from the functional form used and so on and so forth.

- *Simple*: when only one variable Y is connected to an only one variable X; *multiple* when the variables used to explain Y are more than one;
- *Linear*: when Y is expressend using a linear combination of variables $(X_1, \ldots, X_p)$ and parameters $\beta_1, \ldots, \beta_p$, *linearizable*: when originally non-linear models, can be linearized through a transformation (e.g., apply the $\log(\cdot)$ to the Cobb–Douglas production function in the previous example).

For sake of completeness, it is fair to say that there are other classifications that are out of the scope of this course (e.g., spatial, panel models, times series models, non linear models).

Assuming that X is the daily precipitation in a reservoir and Y is the water level of the river, it is likely to assume that there is a relationship of the type $X \to Y$, and not vice versa. If X is the speed of a vehicle and Y the braking distance it makes sense to assume only a one-way relationship $X \to Y$.

If Y is a household's consumption and X is total income, the economic theory proposes various formulations for specifying a functional form $Y = f(X) + \epsilon$. However, for the purposes of a tax investigation, studying a relationship $Y \to X$ and a model $X = f(Y) + \epsilon$ would make sense. Evaluating income through a consumption function would allow verifying the credibility of individual income tax return.

ii) **Estimation Phase**, The estimation phase can be approached as follows. Assume we plan an experiment, determine a random sample of size n, and on each of the statistical units both the phenomenon to be explained and the causes are observed. Thus we have the following data

$$(y_i, x_{i1}, x_{i2}, \ldots, x_{ip}), \text{for } i = 1, \ldots, n,$$

therefore the model specified is the following

$$Y_i = f(x_{i1}, x_{i2}, \ldots, x_{ip}; \beta) + \epsilon_i, \text{ for } i = 1, \ldots, n.$$

However, this model specification has a problem. The variable ϵ_i is not observable, and in the observed data the variable ϵ_i is realized through the number e_i, which can be deduced from the following relation

$$y_i = f(x_{i1}, x_{i2}, \ldots, x_{ip}; \beta) + e_i, \text{ for } i = 1, \ldots, n.$$

Therefore, $e_i = y_i - f(x_{i1}, x_{i2}, \ldots, x_{ip}; \beta)$ are the sample estimate of ϵ_i, and y_i are the observed realizations of the r.v. Y, $i = 1, \ldots, n$.

iii) **Validation Phase** consists of a series of inferential decisions, formalized through hypothesis tests, with the ultimate aim of criticizing and discussing the results obtained. If the validation does not lead to the rejection of the estimated model then this model can be used. Otherwise it has to be reformulated and the information that led to that functional form revised.

A **simple linear regression model** can be expressed using the following form

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n, \tag{2.1}$$

Where β_0 is the intercept and β_1 is the slope. A model is defined to be linear as long as it is linear in the parameters. If $\beta_0 = 0$ the straight line passes though the origin. If $\beta_1 > 0$ ($\beta_1 < 0$), our straight line has positive (negative) slope and as X grows then Y grows (decreases). Finally, if $\beta_1 = 0$ we have a parallel line to the x-axis. Note that the form of the model implies that the only random variable is the response variable Y. The explanatory variable X is usually assumed to be fixed, i.e. measured without error.

For each statistical unit (y_i, x_i), our regression model is based on a **deterministic component**, a function of the variable X ($\beta_0 + \beta_1 x_i$) plus a **stochastic component** (given by the error term ϵ_i). Assuming that $\mathbb{E}(\epsilon_i) = 0, i = 1, \ldots, n$ and that all the x_i are deterministic, the expected value and the variance of Y can be proved to be

$$\mathbb{E}(Y_i) = \mathbb{E}(\beta_0 + \beta_1 x_i + \epsilon_i) = \beta_0 + \beta_1 x_i + \mathbb{E}(\epsilon_i) = \beta_0 + \beta_1 x_i,$$
$$V(Y_i) = V(\beta_0 + \beta_1 x_i + \epsilon_i) = V(\epsilon_i),$$

when the $X's$ are not fixed by the design of the study it is more appropriate to write $\mathbb{E}(Y_i|X_i = x_i)$ in order to make clear that the x-values are themselves random variables but what we are interested in not the distribution of the x- values, only the way Y depends on them. Therefore

$$\mathbb{E}(Y_i|X_i = x_i) = \mathbb{E}(\beta_0 + \beta_1 x_i + \epsilon_i|X_i = x_i) = \beta_0 + \beta_1 x_i + \mathbb{E}(\epsilon_i|X_i = x_i) = \beta_0 + \beta_1 x_i,$$
$$V(Y_i|X_i = x_i) = V(\beta_0 + \beta_1 x_i + \epsilon_i|X_i = x_i) = V(\epsilon_i|X_i = x_i).$$

We are ready to discuss some inferential results, assume to have a sample of size n such that $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ are realizations of X and Y. For the ith statistical unit, we can write $\mathbb{E}(Y_i) = \beta_0 + \beta_1 x_i$, $i = 1, \ldots, n$. Having observed the value of y_i for the r.v. Y_i, the difference $e_i = y_i - (\beta_0 + \beta_1 x_i)$ represents the i^{th} realization of the random variable (r.v.) ϵ_i provided that we know the values for β_0 and β_1. So if the model were a perfect representation of reality for Y, we would observe the exact value for $(\beta_0 + \beta_1 x_i)$, instead we observed y_i. The difference between the observed value for y_i for the r.v. Y_i and the expected value $\beta_0 + \beta_1 x_i$ computed using the linear model when $X_i = x_i$, is defined as the realization e_i (residual(s)) of the r.v. ϵ_i (error(s)).

$$e_i = y_i - (\beta_0 + \beta_1 x_i) \text{ for } i = 1, \ldots, n.$$

For a correct understanding, we must fix the following elements

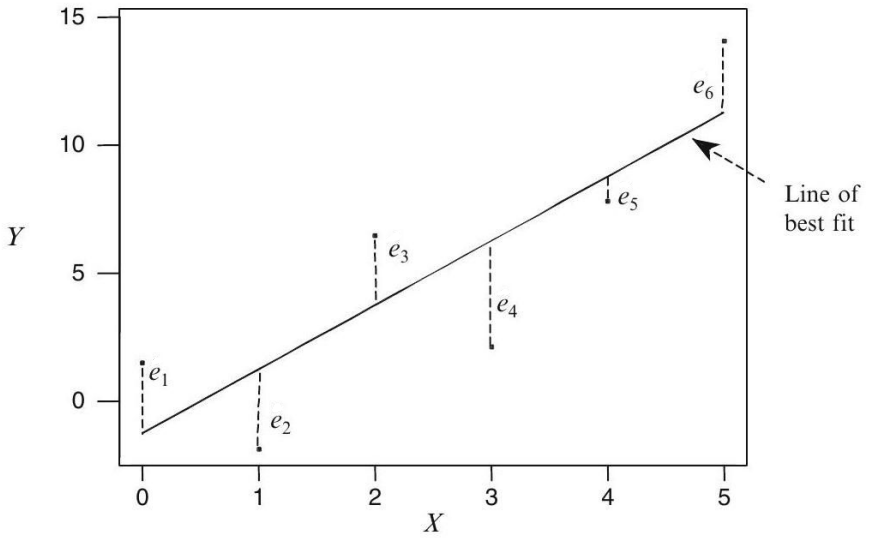
- Model Specification
$$Y = \beta_0 + \beta_1 X + \boldsymbol{\epsilon};$$
- Population
$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \;\; i = 1, \ldots;$$
- Observed Sample
$$y_i = \beta_0 + \beta_1 x_i + e_i, \;\; i = 1, \ldots, n.$$

In practice, we wish to minimize the difference between the actual value of $y\,(y_{\text{i}})$ and the predicted value of y that we will denote with $\hat{y}_i$. This difference is called the residual, e_i, that is,

$$e_i = y_i - \hat{y}_i.$$

Figure **??** shows a hypothetical situation based on six data points. Marked on this plot is a line of best fit, $\hat{y}_i$ along with the residuals.

A very popular method of choosing β_0 and β_1 is called the method of least squares. As the name suggests β_0 and β_1 are chosen to minimize the sum of squared residuals (or residual sum of squares, RSS),



2.1 Estimation and Inference

We can now introduce some assumptions. These conditions are not strictly needed to justify the use of the ordinarily least square (OLS) estimator. They just constitute a simple case in which the small sample properties of the estimator of β_0 and β_1 are easily derived.

- $(A0)$ $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, for $i = 1, \ldots, n$;
- $(A1)$ $\mathbb{E}(\epsilon_i) = 0$, for $i = 1, \ldots, n$;
- $(A2)$ $V(\epsilon_i) = \sigma^2 < +\infty$, for $i = 1, \ldots, n$;
- $(A3)$ $Cov(\epsilon_i, \epsilon_j) = 0$, for $i \neq j, \ldots, n$;
- $(A4)$ X is deterministic, and it is observed for at least two distinct values.

The condition $(A0)$ imposes that the model we are specifying has the same parameters for each observation, i.e. there are no structural breaks in the data, we are also assuming that the model is linear. Condition $(A1)$ states the the expected value of the error terms is zero, on average the regression line should be correct. Assumption $(A2)$ states that the error terms have equal variance, which is referred to as homoskedasticity. Assumption $(A3)$ imposes zero correlation between different error terms. $(A4)$ requires knowledge of the variable X, so we are assuming that the

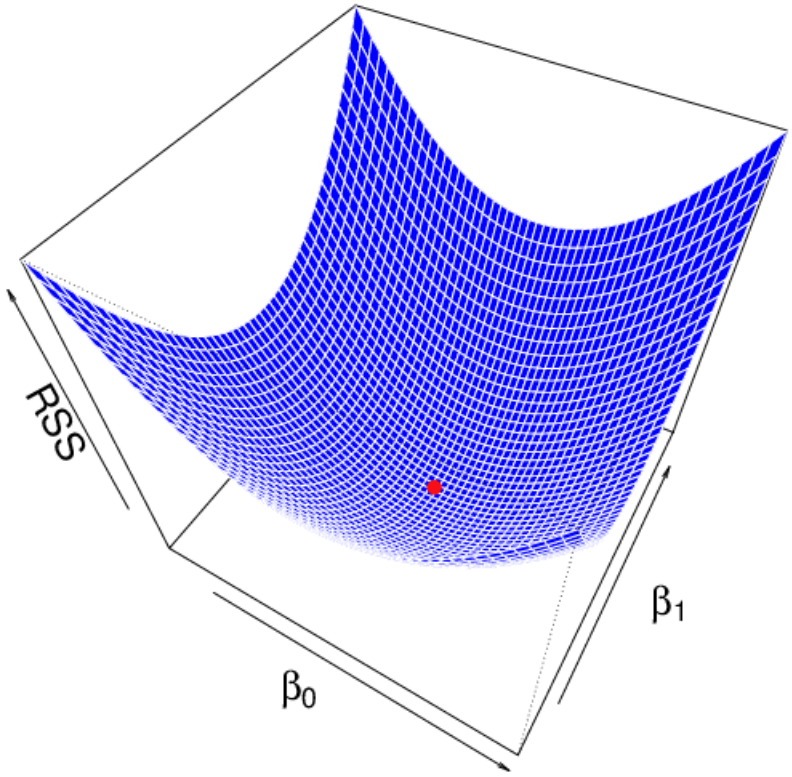


Figure 2.1: 3D plot of the Residual Sum of Squares as a function of the parameters β_0 and β_1, with the red dot representing the minimum of the function, which corresponds to the values that minimize the RSS.

observed sample is the result of a controlled experiment where X is a deterministic variable. However, we have already discussed how the model can be interpreted as a relationship between the conditional mean value of Y and the observed value of X

We are seeking the values of β_0 and β_1 that minimize the residual sum of squares (RSS)

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^{n} (e_i)^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2, \tag{2.2}$$

this is a quadratic expression of the variables (β_0 and β_1). Figure **??** depicts precisely a 3D representation of the objective function expressed in Equation (**??**).

By differentiating respect to β_0 and β_1, we obtain the *normal equations*

$$\begin{cases} 2(-1) \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)) = 0 \\ 2(-1) \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)) x_i = 0 \end{cases}$$

then

$$\begin{cases} \sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)) = 0 \\ \sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))x_i = 0 \end{cases}$$

from which we can obtain the following solutions

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

where

$$\bar{x} = n^{-1}\sum_{i=1}^{n} x_i, \;\; \bar{y} = n^{-1}\sum_{i=1}^{n} y_i, \;\; S_x^2 = n^{-1}\sum_{i=1}^{n}(x_i - \bar{x}), \;\; S_{xy} = n^{-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}),$$

from a computational point of view, the quantities needed to compute the estimates are

$$\sum_{i=1}^{n} x_i \;\; \sum_{i=1}^{n} y_i, \;\; \sum_{i=1}^{n} x_i^2, \;\; \sum_{i=1}^{n} y_i^2, \;\; \sum_{i=1}^{n} x_i y_i,$$

it is useful to recall some alternative formulations

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n}(x_i - \bar{x})y_i - \bar{y}\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n}(x_i - \bar{x})y_i,$$

and the expression of the slope can be written as

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

the estimation of the coefficients δ_i is such that

$$\delta_i = \frac{(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, i = 1, \ldots, n; \;\; \sum_{i=1}^{n}\delta_i = 0; \;\; \sum_{i=1}^{n}(\delta_i)^2 = \sum_{i=1}^{n}\left[\frac{(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]^2 = \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Furthermore, we can rewrite the slope as

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

where

$$b_i = \frac{(y_i - \bar{y})}{(x_i - \bar{x})}; \;\; w_i = \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \; i = 1, \ldots, n; \;\; w_i \geq 0 \;\; \sum_{i=1}^{n} w_i = 1,$$

These formulations show how the OLS estimate of the slope is a weighted average of all the slopes b_i, determined by the straight lines joining the single observed data points (x_i, y_i) with the point $(\bar{x}, \bar{y})$. We note that the weights w_i are an increasing function of $(x_i - \bar{x})$. This means that data points far from the sample mean will have a greater impact in determining the slope. Another way to conceptualize this is by regarding the dispersion of x_i around the mean as equivalent to information. The greater the dispersion among the x_i values, the more information we possess, and consequently, we will have a richer source of information for estimating our parameters. Conversely, when the data cluster closely around the mean, we have less information available, leading to less accurate parameter estimates.

Finally, it is important to discuss what happens when dealing with standardized data points, in this case

$$\frac{\hat{y} - \bar{y}}{S_y} = \beta^* \frac{x_i - \bar{x}}{S_x} \rightarrow (\hat{y} - \bar{y}) = \beta^* \frac{S_y}{S_x}(x_i - \bar{x}), \;\; i = 1, \ldots, n,$$

From which we have

$$\hat{\beta}_1 = \beta^* \frac{S_y}{S_x} \rightarrow \beta^* = \hat{\beta}_1 \frac{S_x}{S_y} = \frac{S_{xy}}{S_x^2} \frac{S_x}{S_y} = r_{xy},$$

Where r_{xy} is the correlation coefficient. Then, using the standardized variables, the intercept reduces the correlation coefficient. So by estimating β^* we are not adding more information than computing the correlation coefficient.

Using the results just discussed, we can prove some important properties about the regression line.

- The regression line is unique, because the minimum of the convex function $RSS(\beta_0, \beta_1)$ is unique;
- The regression line always passes through the point $(\bar{x}, \bar{y})$. From the relation $\hat{\beta}_0 = \bar{y} - \bar{\beta}_1 \bar{x}$, we can deduce that $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$;
- The sample mean of y_i is equal to the $\bar{\hat{y}}_i$. In fact $\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \rightarrow \sum_{i=1}^{n}(y_i - \hat{y}_i) \rightarrow \sum_{i=1}^{n} \hat{y}_i = \sum_{i=1}^{n} y_i$.

Note that the presence of the intercept in the model formulation implies that the sum of the residuals is always zero. Usually, if not for theoretical reasons or specialized knowledge, it is a mistake to omit the intercept from the estimation.

Example 3. *Let's assume to have the following data*

x	y
9,705	3,816
7,267	3,130
8,459	2,955
12,476	4,809
10,296	4,269
8,424	3,291
7,910	2,274
8,879	3,308
11,160	4,340
5,295	1,948
8,421	3,715
12,232	5,340
5,422	2,212
9,900	2,512
12,441	5,277

Table 2.1: Data, Example

we can compute

$$\sum x_i = 138,287; \sum y_i = 53,196; \sum x_i^2 = 1,347,881,959; \sum x_i y_i = 521,674,875; \sum y_i^2 = 205,383,190$$

and

$$\bar{x} = 9,219.133; \bar{y} = 3,549.400; S_{xy} = 2,083,590.5467; S_x^2 = 4,866,377.8489$$

Therefore,

$$\hat{\beta}_1 = \frac{2,083,590.5467}{4,866,377.8489} = 0.4282; \hat{\beta}_0 = (3.546,400) - 0.4282 \times (9,219.133) = -401.2328$$

The linear regression will be

$$\hat{y} = -401.2328 + 0.4282x$$

2.1.1 Gauss Markov Theorem

The estimates $(\hat{\beta}_0, \hat{\beta}_1)$ obtained with the OLS method using observed sample $\{(x_i, y_i), i = 1, \ldots, n\}$ vary as the random sample varies $\{(x_i, Y_i), i = 1, \ldots, n\}$, thus generating a pair of random variables (B_0, B_1), i.e. the corresponding estimators of the parameters. In the regression model the random component is given by ϵ_i, and therefore Y is also a random variable. The estimators are then

$$\begin{cases} B_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ B_0 = \bar{Y} - B_1 \bar{x} \end{cases}$$

In the linear regression model we denote (β_0, β_1) the population parameters, with $(\hat{\beta}_0, \hat{\beta}_1)$ the estimates obtained from OLS (= numbers), with (B_0, B_1) are the estimators (=r.v.) generated from the random sampling process. We are now ready to discuss one of the most important results in Inferential Statistics

Theorem 2. *(Gauss-Markov Theorem). Under the classical assumptions of the simple regression model, OLS estimators (B_0, B_1) for the parameters (β_0, β_1) are linear, unbiased, and the most efficient in the class of linear and unbiased estimators (BLUE) (A proof based on the multivariate version of this theorem can be found in the following section.)*

To employ the least squares estimators for making meaningful inferences, we need to discuss their statistical properties. We have derived the expression of B_1 and B_0, it can be showed that the least squares estimators are unbiased

$$\begin{aligned} \mathbb{E}(B_1) &= \mathbb{E}\left[\frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2}\right] = \frac{\sum_{i=1}^n (x_i - \bar{x})\mathbb{E}(Y_i)}{\sum_{i=1}^n (x_i - \bar{x})^2}, \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})(\beta_0 + \beta_1 x_i)}{\sum_{i=1}^n (x_i - \bar{x})^2}, \\ &= \beta_0 \frac{\sum_{i=1}^n (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} + \beta_1 \frac{\sum_{i=1}^n (x_i - \bar{x}) x_i}{\sum_{i=1}^n (x_i - \bar{x})^2}, \\ &= 0 + \beta_1 \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \beta_1. \end{aligned}$$

To find $V(B_1)$ we need to recall that $Y_1, Y_2, Y_3, \ldots, Y_n$ are independent random variables, therefore

$$\begin{aligned} V(B_1) = V\left[\frac{\sum_{i=1}^n (x_i - \bar{x})Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}\right] &= \frac{1}{S_{xx}^2} \sum_{i=1}^n V\left[(x_i - \bar{x})Y_i\right], \\ &= \frac{1}{S_{xx}^2} \sum_{i=1}^n (x_i - \bar{x})^2 V(Y_i), \end{aligned}$$

where $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$, recalling that $V(Y_i) = \sigma^2$, $i = 1, \ldots, n$

$$V(B_1) = \sigma^2 / S_{xx},$$

to study the properties of B_0 we should first discuss some results

$$\mathbb{E}(\bar{Y}) = \mathbb{E}(n^{-1}\sum(Y_i)) = n^{-1}\sum \mathbb{E}(\beta_0 + \beta_1 x_i + \epsilon_i) = \beta_0 + \beta_1 \bar{x},$$

it can be proved that the covariance between $\bar{Y}$ and B_1 is zero

$$Cov(\bar{Y}, B_1) = 0,$$

we have that

$$\mathbb{E}(B_0) = \mathbb{E}(\bar{Y}) - \mathbb{E}(B_1)\bar{x} = \beta_0 + \beta_1\bar{x} - \beta_1\bar{x} = \beta_0,$$

as we know the expressions for $V(\bar{Y})$, $V(\hat{\beta}_1)$ and $Cov(\bar{Y}, \hat{\beta}_1)$, we can derive the variance of B_0

$$V(B_0) = \frac{\sigma^2}{n}\left[1 + \frac{n(\bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right],$$

finally, we can prove that B_0 and B_1 are correlated

$$Cov(B_0, B_1) = \frac{-\bar{x}\sigma^2}{S_{xx}}.$$

Unfortunately, the value of σ^2 is unknown, so we have to use our sample information to find a reasonable estimator. Because we are using $\hat{Y}$ to estimate $\mathbb{E}(Y_i)$ it seems natural to base an estimate of σ^2 upon $RSS = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$. Indeed it can be showed that

$$S^2 = \left(\frac{1}{n-2}\right)\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \left(\frac{1}{n-2}\right)RSS,$$

is an unbiased estimator for σ^2. We note how the variability of the estimates decreases as the quantity $\sum_{i=1}^n (x_i - \bar{x})^2$ increases, and therefore as the variance of the sample increases, the variability must be seen as a source of information and, specifically, only by having an adequate variability for the observations will it be possible to check if this variability has an effect on Y.

It can be also proved that $B_0 \xrightarrow{p} \beta_0$ and $B_1 \xrightarrow{p} \beta_1$ meaning that both B_0 and B_1 are consistent estimators. If we assume that ϵ_i are independent, then S^2 is a consistent estimator for σ^2. Finally B_0 and B_1 are also asymptotically normal.

Plugging in the expression of S^2 in $V(B_0)$ and $V(B_1)$ we obtain the estimators for the variances, the standard errors

$$se^2(B_0) = \frac{S^2}{n}\left[1 + \frac{n(\bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right],$$
$$se^2(B_1) = \frac{S^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Where the square root of the quantities just introduced constitutes the standard errors of B_0 and B_1, respectively. The standard error is a measure of the variability or uncertainty associated with the estimated coefficients of the regression equation. It provides an indication of how reliable or precise the estimated coefficients are in representing the true relationships between the independent and dependent variables. The distinction between the variance and the standard errors of B_0 and B_1 lies in the fact that the variance cannot be directly calculated, as it relies on σ^2 (the variance of the errors).

2.1.2 Maximum likelihood estimators for β_1 and β_0

Assuming that the r.v. are i.i.d. (independent and identically distributed) $\epsilon_i \sim N(0, \sigma^2)$, then the observed sample $(y_1, \ldots, y_n)$ is the realization of the random sample $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ whose components are distributed as $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$. Therefore the likelihood function can be written

$$L(\beta_0, \beta_1, \sigma^2|\mathbf{y}) = (2\pi\sigma^2)^{-n/2} \exp\left[- \frac{\sum_i (y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2} \right],$$

and the log-likelihood function is

$$\ell(\beta_0, \beta_1, \sigma^2) \propto -(n/2)\log(\sigma^2) - \frac{\sum_i (y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2},$$

where $\propto$ means proportional to, taking the derivatives with respect to β_0 and β_1 we end up with the Maximum Likelihood estimators(MLE)

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

We note that the estimates and the corresponding MLE estimators coincide with the expressions we found for the OLS. However, the estimate of the variance is biased.

Therefore MLE enjoy all the beautiful properties we have discussed, also for MLE estimators the Gauss Markov theorem still holds, but because we are using maximum likelihood to obtain those expressions, our estimators are also sufficient and efficient among all unbiased estimators (not just linear and unbiased). In fact, it can be proved that the MLE estimators reach the Cramer-Rao lower bound. Furthermore for any value of n we have that these estimators are also normally distributed. In the case of OLS we had to impose $n \to \infty$ to obtain an asymptotically normal distribution.

Under this domain, the normality hypothesis plays a crucial role on the efficiency of these estimators. Therefore, the hypothesis of normality on the residuals will have to be carefully assessed.

2.1.3 Inference about the paramaters β_i

Now that we have estimated the model using the ordinary least squares (OLS) method and maximum likelihood, we want to evaluate the significance of the parameters and then move on to the validation phase. Using the assumption of error normality, the random variable $(B_0, B_1)^\top$ is essentially a bivariate normal random variable with mean vector $\boldsymbol{\beta} = (\beta_0, \beta_1)$ and variance-covariance matrix given by $\sigma^2\boldsymbol{V}$. Where

$$\mathrm{V}\left(\begin{bmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{bmatrix}\right) = \sigma^2\boldsymbol{V} = \sigma^2\left(\begin{bmatrix}\frac{1}{n}[1+\frac{n(\bar{x}^2)}{\sum_{i=1}^n(x_i-\bar{x})^2}] & -\frac{\bar{x}}{\sum_{i=1}^n(x_i-\bar{x})^2}\\ -\frac{\bar{x}}{\sum_{i=1}^n(x_i-\bar{x})^2} & \frac{1}{\sum_{i=1}^n(x_i-\bar{x})^2}\end{bmatrix}\right),$$

where: σ^2 is the variance of the errors ϵ_i, $\bar{x}$ is the mean of the x_i values, the denominators represent the sum of the squared deviations of the predictor variable from its mean.

Suppose we have fitted a linear model and we want to test

$$\begin{aligned} H_0 :&\beta_0 = b_0 \text{ vs } H_1 : \beta_0 \neq b_0\\ H_0 :&\beta_1 = b_1 \text{ vs } H_1 : \beta_1 \neq b_1, \end{aligned}$$

where b_0 and b_1 are the values we want to test. In order to test this set of hypotheses, we derive a test statistic. Recall from the previous paragraph, the unbiased estimator for σ^2 is S^2.

Given that Z is a standardized normal random variable and χ^2 is an independent chi-squared random variable with v degrees of freedom, the ratio

$$T = \frac{Z}{\sqrt{\frac{\chi^2_\nu}{v}}},$$

follows a T-distribution with v degrees of freedom. If the hypothesized model is true, then:

$$\frac{\frac{(B_i-\beta_i)}{\sqrt{V(B_i)}}}{\sqrt{\frac{S^2}{\sigma^2}}} \sim T_{n-2} \text{ for } i = 0, 1.$$

We note that these random variables do not depend on σ^2 as it is eliminated in the ratio. Thus, the following testing scheme applies:

$$\begin{aligned} H_0 :&\beta_0 = b_0 \text{ vs } H_1 : \beta_0 \neq b_0 \qquad \text{we reject } H_0 \text{ if } |T_0| \geq t_{\alpha/2;n-1}\\ H_0 :&\beta_1 = b_1 \text{ vs } H_1 : \beta_1 \neq b_1 \qquad \text{we reject } H_0 \text{ if } |T_1| \geq t_{\alpha/2;n-1} \end{aligned}$$

where

$$T_0 = \frac{\hat{\beta}_0 - b_0}{se(B_0)}, \quad T_1 = \frac{\hat{\beta}_1 - b_1}{se(B_1)}.$$

Similarly, confidence intervals (CI) for the individual parameters can be constructed:

$$\text{Confidence interval for } \beta_0 : \hat{\beta}_0 - t_{\alpha/2;n-2} se(B_0) \leq \beta_0 \leq \hat{\beta}_0 + t_{\alpha/2;n-2} se(B_0)$$
$$\text{Confidence interval for } \beta_1 : \hat{\beta}_1 - t_{\alpha/2;n-2} se(B_1) \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2;n-2} se(B_1).$$

The value $t_{\alpha/2,n-1}$ from the Student's t-distribution with $n-1$ degrees of freedom is defined as the value such that the area to the left of this value is $\alpha/2$ and the area to the right of its symmetric value (i.e., the negative of this value) is also $\alpha/2$.

In simpler terms, for a two-tailed test:

- The area to the left of $t_{\alpha/2,n-1}$ is $\alpha/2$.
- The area to the right of $-t_{\alpha/2,n-1}$ (the negative counterpart) is $\alpha/2$.

These quantiles are commonly used in hypothesis testing and the construction of confidence intervals when the population variance is unknown and the sample size is small. For a two-tailed test at the 5% significance level, where $\alpha = 0.05$ and $\alpha/2 = 0.025$, the quantiles $t_{0.025,n-1}$ and $t_{0.975,n-1}$ provide the critical values for the test or the bounds of a 95% confidence interval.

2.2 Goodness of fit

When estimating parameters for a regression model, it is always advisable to propose an indicator capable of summarizing the explanatory power of the model in relation to the sample data. This can be done by calculating the R^2 index, which can be derived from the decomposition of the total deviance. Let's see how.

$$\text{Dev}(y) = \text{Dev}(\hat{y}) + \text{Dev}(e),$$

where

- $\text{Dev}(y) = \sum_{i=1}^{n}(y_i - \bar{y})^2$=Total Deviance=TSS;
- $\text{Dev}(\hat{y}) = \sum_{i=1}^{n}(y_i - \bar{\hat{y}})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$=Regression Deviance=SSR;
- $\text{Dev}(e) = \sum_{i=1}^{n}(e_i - 0)^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$=Residual Deviance =RSS.

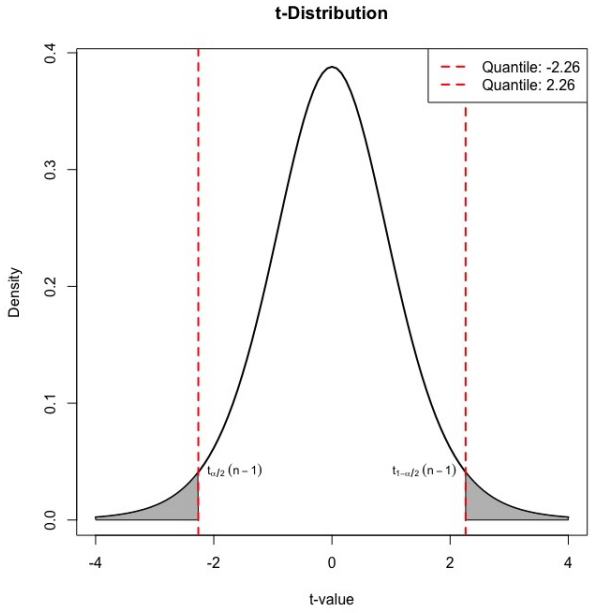


Figure 2.2: Probability density function (PDF) of a t-distribution with $n-1$ degrees of freedom, representative of a sample size of n. The symmetric curve highlights the central tendency and variability of the t-distribution. Two critical quantiles, $t_{\alpha/2,n-1}$ and $t_{1-\alpha/2,n-1}$, are marked by red dashed lines, indicating the tail regions of the distribution for a significance level of α. The shaded areas under the curve on either side represent the cumulative probabilities in the tails beyond these critical quantiles. For the given example, these quantiles are approximately -2.26 and 2.26 respectively, providing boundaries for hypothesis testing at the 5% significance level. The area between these two quantiles contains 95% of the distribution's total area, leaving 2.5% in each tail.

The proposed decomposition exists only when including the intercept. Now, if the variability of Y is largely explained by the regression line, the regression deviance (SSR) will be high relative to the total deviance (TSS), and consequently, the residual deviance (RSS) will be low. It follows that a measure of the goodness of the regression line is

$$R^2 = \frac{\text{Dev}(\hat{y})}{\text{Dev}(y)} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

Equivalently, thanks to the decomposition of the deviance, we can write

$$R^2 = 1 - \frac{\text{Dev}(e)}{\text{Dev}(y)} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

The presented index represents the proportion of the total variance explained by the regression variance. Therefore, $0 \leq R^2 \leq 1$.

It is possible to prove that in the case of simple linear regression, the coefficient R^2 is the square of the correlation index calculated between the y_i and x_i.

$$R^2 = 1 - \frac{(n-2)s^2}{ns_y^2} = 1 - \frac{n[s_y^2 - (\hat{\beta}_1)^2 s_x^2]}{ns_y^2} = (\hat{\beta}_1)^2 \frac{s_x^2}{s_y^2} = \left(\frac{s_{xy}}{s_x^2}\right)^2 \frac{s_x^2}{s_y^2} = \left(\frac{s_{xy}}{s_x s_y}\right)^2 = (r_{xy})^2.$$

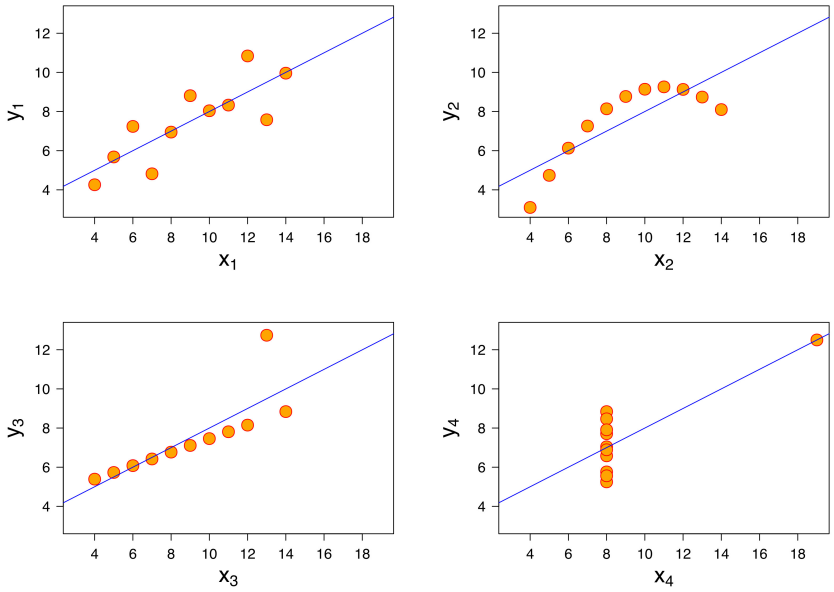


Figure 2.3: Anscombe's quartet is a famous statistical example that consists of four datasets with nearly identical descriptive statistics but vastly different visual representations each having an equal R^2 (0.67). The quartet was created by the statistician Francis Anscombe in 1973.

Despite the R-squared index being a powerful tool for assessing the goodness of fit, in real-world applications, everything becomes more complex. Therefore, relying solely on summary statistics can be misleading. An example of this is the Anscombe's quartet in Figure **??**. In the Figure **??**, four datasets are plotted, each having an equal coefficient of determination R^2 (0.67). It is clear, therefore, that before estimating any model, every analyst must explore the data, ask questions, and test hypotheses.

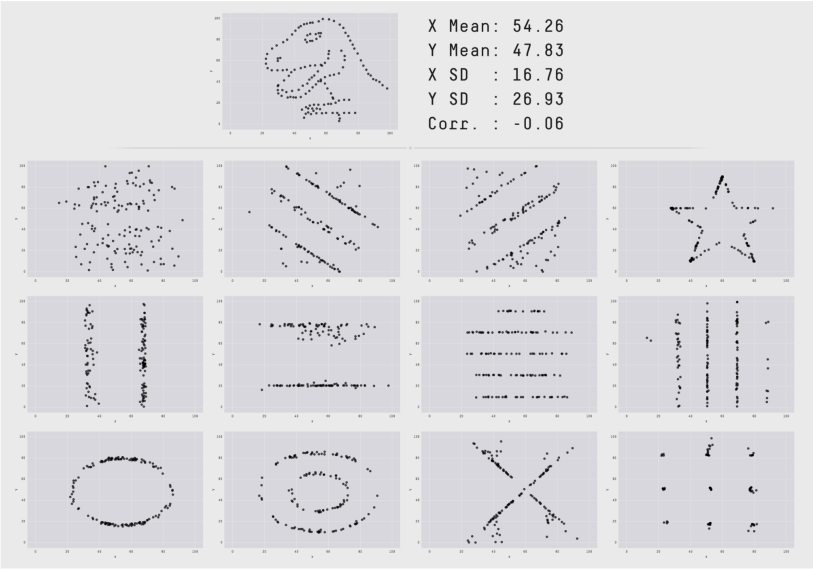


Figure 2.4: Another example of why we should never rely solely on summary statistics is that the datasets all have the same mean, standard deviation, and correlation.

Example 4. *(Easy way to compute $\hat{\beta}_1$ and $\hat{\beta}_0$)*

A rocket motor is manufactured by bonding an igniter propellant and a sustainer propellant together inside a metal housing. The shear strength of the bond between the two types of propellant is a critical quality characteristic. It is hypothesized that the shear strength may be influenced by the age (in weeks) of the sustainer propellant batch. A set of twenty observations concerning shear strength and the age of the corresponding propellant batch have been gathered and are presented in Table **??**. *Figure* **??** *depicts a scatter diagram, indicating a significant statistical relationship between shear strength and propellant age. The preliminary assumption of the linear model $y = \beta_0 + \beta_1 x + \epsilon$ seems justifiable.*

Table 2.2: Observations on Shear Strength and Age of Propellant

X Age of Propellant (weeks)	Y Shear Strength (psi)
15.50	2158.70
23.75	1678.15
8.00	2316.00
17.00	2061.30
5.50	2207.50
19.00	1708.30
24.00	1784.70
2.50	2575.00
7.50	2357.90
11.00	2256.70
13.00	2165.20
3.75	2399.55
25.00	1779.80
9.75	2336.75
22.00	1765.30
18.00	2053.50
6.00	2414.40
12.50	2200.50
2.00	2654.20
21.50	1753.70

In simple linear regression, we must calculate both $\hat{\beta}_0$ and $\hat{\beta}_1$. While the former is relatively straightforward to compute, the latter can become cumbersome, especially with a large number of data points. In this exercise we propose a easier method to compute this quantity

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

With a bit of algebra, we can rewrite this expression, starting from the S_{xy} we have

$$\begin{aligned} S_{xy} = \sum_{i=1}^{n} y_i(x_i - \bar{x}) &= \sum_{i=1}^{n} x_i y_i - \bar{x} \sum_{i=1}^{n} y_i \\ &= \sum_{i=1}^{n} y_i x_i - \bar{x} \frac{n}{n} \sum_{i=1}^{n} y_i \\ &= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} \\ &= \sum_{i=1}^{n} x_i y_i - n \frac{\sum_{i=1}^{n} x_i}{n} \frac{\sum_{i=1}^{n} y_i}{n} \\ &= \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}. \end{aligned}$$

While $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$ can be rewritten in this way

$$\begin{aligned} S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 &= \sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\ &= \sum_{i=1}^{n} x_i^2 - 2\bar{x} \sum_{i=1}^{n} x_i + n\bar{x}^2 \\ &= \sum_{i=1}^{n} x_i^2 - 2\frac{\sum_{i=1}^{n} x_i}{n} \sum_{i=1}^{n} x_i + n\left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2 \\ &= \sum_{i=1}^{n} x_i^2 - 2\frac{\sum_{i=1}^{n} x_i}{n} \sum_{i=1}^{n} x_i + n\frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n^2} \\ &= \sum_{i=1}^{n} x_i^2 - 2\frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} + \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} \\ &= \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}. \end{aligned}$$

Now we are ready to use these formulas in the exercise. Let's see how

	x_i	y_i	x_i^2	$y_i x_i$
	15.50	2158.70	240.25	33459.85
	23.75	1678.15	564.06	39856.06
	8.00	2316.00	64.00	18528.00
	17.00	2061.30	289.00	35042.10
	5.50	2207.50	30.25	12141.25
	19.00	1708.30	361.00	32457.70
	24.00	1784.70	576.00	42832.80
	2.50	2575.00	6.25	6437.50
	7.50	2357.90	56.25	17684.25
	11.00	2256.70	121.00	24823.70
	13.00	2165.20	169.00	28147.60
	3.75	2399.55	14.06	8998.31
	25.00	1779.80	625.00	44495.00
	9.75	2336.75	95.06	22783.31
	22.00	1765.30	484.00	38836.60
	18.00	2053.50	324.00	36963.00
	6.00	2414.40	36.00	14486.40
	12.50	2200.50	156.25	27506.25
	2.00	2654.20	4.00	5308.40
	21.50	1753.70	462.25	37704.55
Column sum	267.250	42627.150	4677.688	528492.637

Table 2.3: Calculations needed for the estimation of the regression line.

We have that $\bar{x} = 267.250/20 = 13.3625$, $\bar{y} = 42627.150/20 = 2131.358$ *and* $\sum_{i=1}^{n} x_i y_i = 528492.637$, *then*

$$S_{xy} = 528492.637 - \frac{(267.250)(42627.150)}{20} = -41112.65,$$

while, knowing that $\sum_{i=1}^{n} x_i^2 = 4677.688$ *and* $\left(\sum_{i=1}^{n} x_i\right)^2 = 267.250^2 = 71422.56$ *we have*

$$S_{xx} = 4677.688 - \frac{267.250^2}{20} = 1106.56.$$

Finally

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{-41112.65}{1106.56} = -37.15,$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 2131.3575 - (-37.15 \times 13.3625) = 2627.82$$

The least-squares fit is

$$\hat{y} = 2627.82 - 37.15x$$

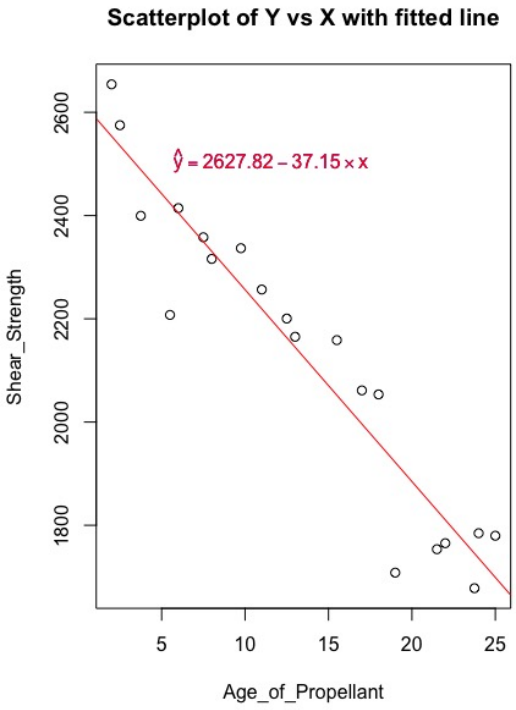


Figure 2.5: Scatterplot of Shear Strength vs Age of Propellant with fitted regression line.

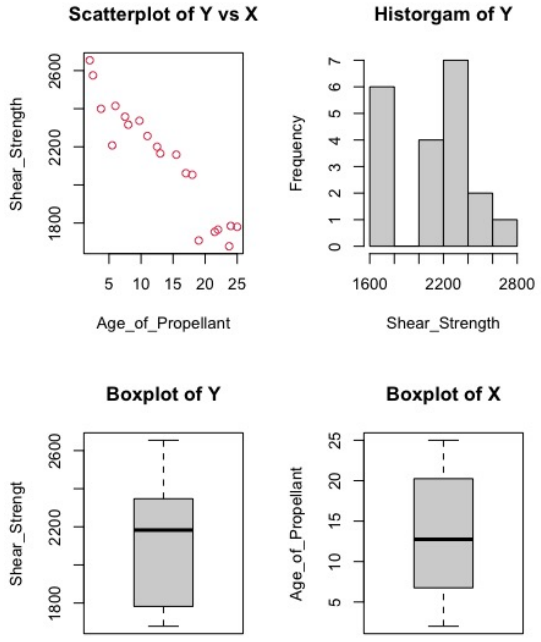


Figure 2.6: Exploratory analysis of propellant data, illustrating the relationship between the shear strength and the age of the propellant.

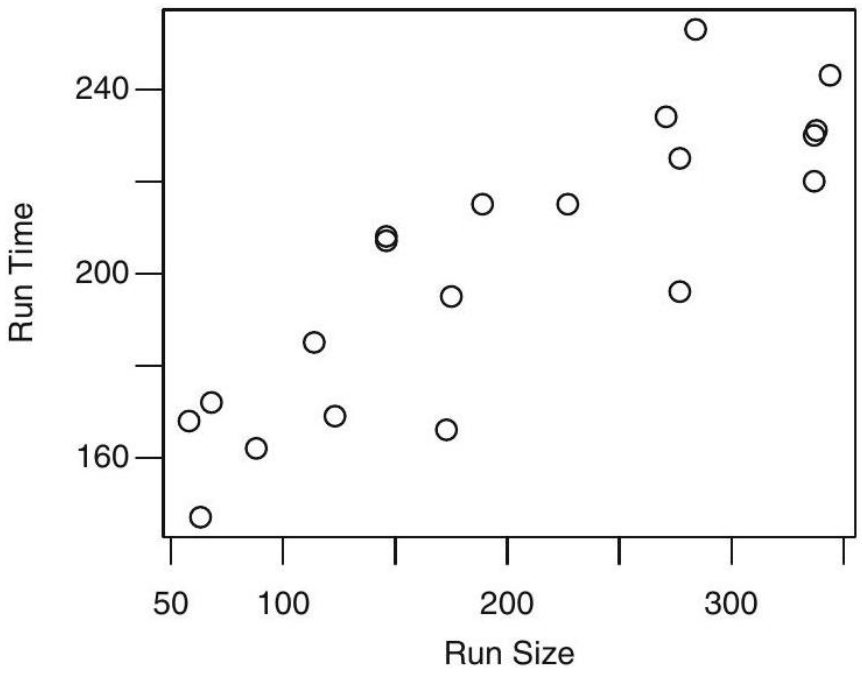


Figure 2.7: Scatterplot of the production data

2.3 Regression Output from R

In this example, we consider some data taken from Foster, Stine and Waterman (1997, pages 191-199). The original data are in the form of the time taken (in minutes) for a production run, Y and the number of items produced, X, for 20 randomly selected orders as supervised by three managers. We consider the data for one of the managers (see Table and Figure **??**). We wish to develop an equation to model the relationship between Y the run time, and X, the run size.

Case	Run time	Run size	Case	Run time	Run size
1	195	175	11	220	337
2	215	189	12	168	58
3	243	344	13	207	146
4	162	88	14	225	277
5	185	114	15	169	123
6	231	338	16	215	227
7	234	271	17	147	63
8	166	173	18	230	337
9	253	284	19	208	146
10	196	277	20	172	68

Figure **??** shows a scatter plot of the production data. The least squares estimates for the production data were calculated using R, giving the following results:

```
Coefficients:
             Estimate Std. Error t value  Pr(>|t|)
(Intercept)  149.74770    8.32815   17.98  6.00e-13 ***
RunSize        0.25924    0.03714    6.98  1.61e-06 ***
---
Signif. codes:0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.25 on 18 degrees of freedom
Multiple R-Squared: 0.7302,      Adjusted R-squared: 0.7152
F-statistic: 48.72 on 1 and 18 DF, p-value: 1.615e-06
```

Where the intercept is $\hat{\beta}_0 = 149.74$ and the slope $\hat{\beta}_1 = 0.26$. While the standard error of the intercept and of the slope are $se(\hat{\beta}_0) = 8.33$ and $se(\hat{\beta}_1) = 0.04$. **t value**: The t-value is calculated as the estimated coefficient divided by its standard error. It measures the significance of the coefficient. $\mathbf{Pr}(> |\mathbf{t}|)$: The p-value associated with the t-value. It indicates the statistical significance of the coefficient. Generally, a p-value below 0.05 suggests that the coefficient is statistically significant. **Residual standard error:** This value represents the standard deviation of the residuals, which measures the average distance between the observed values and the predicted values by the model. A lower residual standard error indicates a better fit of the model to the data. **R-squared:** This statistic, denoted as R-squared or coefficient of determination, indicates the proportion of the total variance in the dependent variable explained by the regression model. It ranges from 0 to 1, where 0 indicates no explanatory power, and 1 represents a perfect fit. **Adjusted R-squared:** This is a modified version of R-squared that takes into account the number of predictors in the model. It penalizes the addition of unnecessary predictors, providing a more reliable measure of model fit. **F-statistic:** The F-statistic tests the overall significance of the regression model. It assesses whether at least one of the independent variables has a significant effect on the dependent variable. The F-statistic is accompanied by its associated p-value.**Significance codes:** This section provides asterisks () next to the p-values in the coefficients table to indicate the level of significance. For example, a p-value less than 0.001 is often represented as "***".

2.4 Exercises

1. Invent examples of data analytic problems, one each for the three major aims of modelling:

- prediction,
- explanation,
- causal inference.

By "invent" it is meant that you describe the variables, what scientists are interested in, and the background of the situation as far as necessary to understand to which of the three aims you classify the example. You do not need to make up the datasets.

2. Express the sum of squares of the errors in terms of the three unknown parameters $\beta_0, \beta_1, \beta_2$, and then obtain the normal equations by differentiation with respect to each parameter.

3. Write down the matrix $\mathbf{X}$ and the vectors $\boldsymbol{\beta}$ and $\mathbf{Y}$ and verify that the equations obtained in the previous question are of the form $\mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^{\mathrm{T}}\mathbf{Y}$.

4. Suppose that we fit the straight-line regression model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$ but suppose that the rensponse is affected by a second variable x_2 such that the true regression function is

$$\mathbb{E}(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

a) Is the OLS estimator of the slope in the original simple linear regression model unbiased ?
b) Show the bias in B_1.

5. Consider the simple linear regression model

$$y = \beta_0 + \beta_1 x + \epsilon,$$

where the intercept is known
a) Find the OLS estimator of β_1 for this model. Does this answer seem reasonable ?
b) What is the variance of the slope $(\hat{\beta}_1)$ fo the OLS estimator found in part a),
c) Find a $100(1-\alpha)\%$ CI for β_1. Is this interval narrower than the estimator for the case where both slope and intercept are unknown?

6. Convert the following linear relationships into linear relationships by making transformations and defining new variables

a) $y = a/(b + cx)$,
b) $y = ae^{-bx}$,
c) $y = ab^x$,
d) $y = x/(a + bx)$,
e) $y = (1/1 + e^{bx})$.
where a, b, c are constants.

7. Suppose that the model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, the errors have mean zero and are independent, but $Var(\epsilon_i) = \rho_i^2 \sigma^2$, where ρ_i are known constants, so the errors do not have equal variance. this situation arises when the y_i are averages of several observations at x_i; in this case, if y_i is an average of n_i independent observations, $\rho_i^2 = 1/n_i$ (why?). Because the variances are not equal, the theory developed in this chapter does not apply; intuitively, it seems that the observations with large variability should influence the estimates of β_0 and β_1 less than the observations with small variability.

   The problem may be transformed as follows

   $$\rho_i^{-1} y_i = \rho_i^{-1}\beta_0 + \rho_i^{-1}\beta_1 x_i + \rho_i^{-1}\epsilon_i,$$

   or

   $$z_i = u_i\beta_0 + v_i\beta_1 + \delta_i,$$

   where $u_i = \rho_i^{-1}$, $v_i = \rho_i^{-1} x_i$, $\delta_i = \rho_i^{-1}\epsilon_i$,

   a) Show that the new model satisfies the assumptions of the standard statistical model,
   b) Find the least squares estimates for β_0 and β_1,
   c) Show that performing the least squares analysis on the new model, as was done in part b), is equivalent to minimizing.

   $$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \rho_i^{-2}.$$

   This is a weighted least squares criterion; the observations with large variances are weighted less.

8. Suppose that a line is fit by the method of least squares to n points, that the standard statistical model holds, and that we want to estimate the line at a new point, x_0. Denoting the value on the line by μ_0, the estimate is

   $$\hat{\mu}_o = \hat{\beta}_0 + \hat{\beta}_1 x_0.$$

   a) Derive the expression for the variance of $\hat{\mu}_0$,
   b) Sketch the standard deviation of $\hat{\mu}_0$ as a function of $x_0 - \bar{x}$. The shape of the curve should be intuitively plausible,
   c) Derive a 95% confidence interval for $\mu_0 = \beta_0 + \beta_1 x_0$ under the assumption of normality.

Chapter 3

Multiple Linear Regression

In general let p explanatory variables labelled $x_1, \ldots, x_p$. We can consider the following model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i \quad i = 1, \ldots, n, \tag{3.1}$$

note that in the following expression we have p explanatory variable and $p+1$ parameters (p plus the intercept). The model expressed in (**??**) can be still considered a linear model as its predictor is linear in unknown parameters $\beta_1 \ldots, \beta_p$.

The formulation introduced in (**??**) can be generalized using the matrix notation. Let $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^\top$ its realizations $\mathbf{y} = (y_1, y_2, \ldots, y_n)^\top$. The explainatory variables are contained in the matrix $\mathbf{X}$ of dimension $(n \times (p+1))$. Finally, we denote with $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^\top$ the vector containing the error terms. The parameters are $p+1$ and are contained in the vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$. **The multiple linear model** in matrix notation takes the following form

$$\mathbf{Y} = \overbrace{\mathbf{X}\boldsymbol{\beta}}^{Deterministic} + \overbrace{\boldsymbol{\epsilon}}^{Stochastic} \tag{3.2}$$

$$\text{where } \underset{n\times 1}{\mathbf{Y}} = \begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_n \end{bmatrix}, \underset{n\times(p+1)}{\mathbf{X}} = \begin{bmatrix} 1 & x_{11} & x_{12} \ldots x_{1p} \\ 1 & x_{21} & x_{22} \ldots x_{2p} \\ & & . \\ & & . \\ & & . \\ 1 & x_{n1} & x_{n2} \ldots x_{np} \end{bmatrix}, \underset{(p+1)\times 1}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ . \\ \beta_p \end{bmatrix}, \underset{n\times 1}{\boldsymbol{\epsilon}} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ . \\ \epsilon_n \end{bmatrix}$$

for the $i-th$ statistical unit we have

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} +, \ldots, + \beta_p x_{ip} + \epsilon_i, \;\; i = 1, 2, \ldots, n,$$

after having collected a sample, we have the observed values y_i for the r.v. Y and $\{x_{ij}, i = 1, 2, \ldots, n; j = 1, 2, \ldots, p\}$ for the explainatory variables $X_j, j = 1, 2, \ldots, p$. Therefore, we have

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \tag{3.3}$$

for the $i - th$ statistical unit we have

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + e_i \;\; i = 1, \ldots, n.$$

As already discussed in the previous section, the vector $\mathbf{e} = (e_1, \ldots, e_n)^\top$ contains the realizations of the r.v. $\boldsymbol{\epsilon}$ (Stochastic error) which in general is not observed. The realizations of ϵ_i[1] can be determined once the vector of parameters $\boldsymbol{\beta}$ has been estimated and then $\hat{\boldsymbol{\beta}}$ is obtained

$$\mathbf{e} = \mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}.$$

3.1 Estimation

As for the simple linear model case, here too we have hypotheses. Whose can be viewed as a matrix generalization of the ones discussed for the simple linear model.

- (A1) $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$,
- (A2) $\mathbb{E}(\boldsymbol{\epsilon}) = \mathbf{0}$,
- (A3) $Var(\boldsymbol{\epsilon}) = \mathbb{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top) = \sigma^2 \boldsymbol{I}_n$,
- (A4) $\mathbf{X}$ is a deterministic matrix and is full rank, $rank(\mathbf{X}) = p + 1$.

In other words we are assuming i) **Linearity of the model**: the relationship between the dependent variable $\mathbf{Y}$ and the independent variables $\mathbf{X}$ is linear and constant for all observations, this means that there are no structural breaks in the data, an example of structural break is given by the third plot in Figure **??**. ii) **Zero mean of errors**: the errors $\boldsymbol{\epsilon}$ have a zero mean, that is, $\mathbb{E}(\boldsymbol{\epsilon}) = \mathbf{0}$. This assumption means that on average, the model is correct, and any deviations from the true values are random and cancel each other out. iii) **Homoscedasticity**: the errors have constant variance or are homoscedastic, that is, $V(\boldsymbol{\epsilon}) = \mathbb{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top) = \sigma^2 \boldsymbol{I}_n$, where σ^2 is the constant

[1]The distinction between error terms ϵ_i and residuals e_i is important. Error terms are unobservable, and distributional assumptions about them are necessary to derive sampling properties of estimators for $\boldsymbol{\beta}$. The residuals are obtained after estimation, and their value depend upon the estimated values for $\boldsymbol{\beta}$ and therefore depend upon the sample and estimation method.

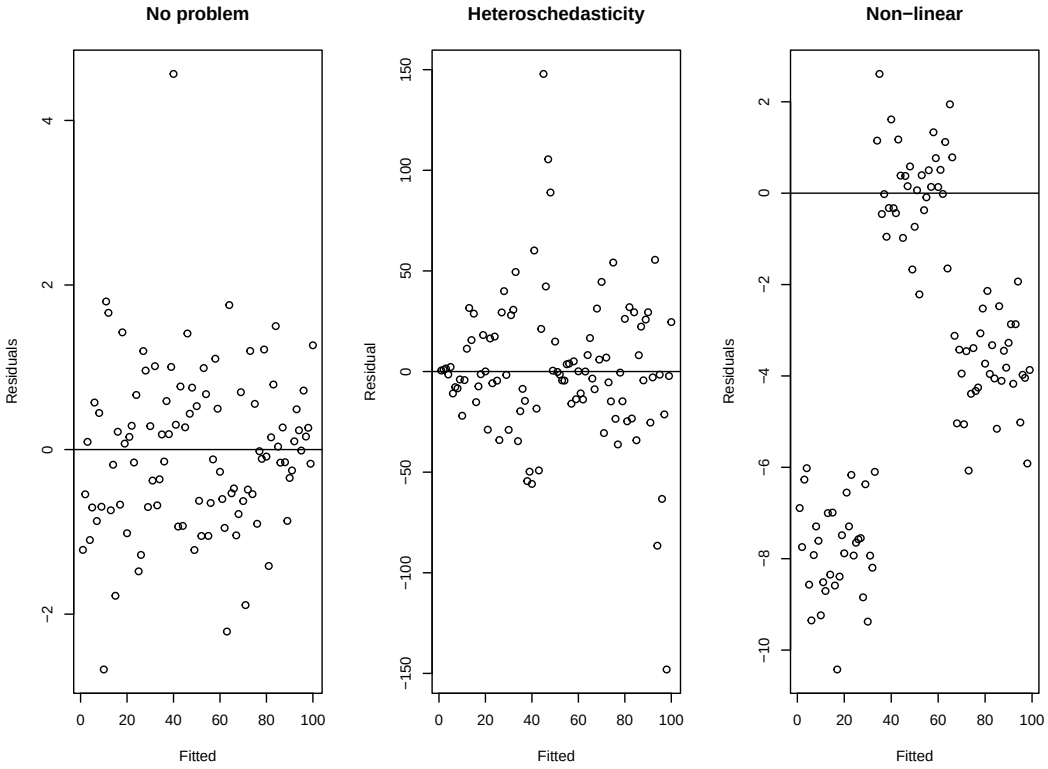


Figure 3.1: Examples of violation of the OLS assumptions. Residuals vs Fitted Values Plot obtained after estimating a multiple linear regression model. The first plot depicts residuals with no issues, as they are centered around zero and exhibit homoscedasticity. The second plot illustrates an instance of heteroscedastic residuals, which remain centered around zero, but their dispersion is no longer constant. The final plot, on the other hand, demonstrates an example of structural break. In practical terms, especially when dealing with economic data, it is possible for the estimated beta value to vary across different data points.

variance of the errors and $\boldsymbol{I}_n$ is the identity matrix of size n, an example is given by the second plot in Figure **??**. iv) **No multicollinearity**: The fourth assumption of OLS is that there is no perfect multicollinearity among the independent variables. This means that the independent variables in $\mathbf{X}$ are not linearly dependent on each other, as this would lead to problems with the estimation of the regression coefficients. Further will be discussed in Section **??**.

We wish to find the vector of least-squares estimators $\hat{\boldsymbol{\beta}}$, that minimizes

$$\begin{aligned} RSS(\boldsymbol{\beta}) = S(\boldsymbol{\beta}) &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2, \end{aligned} \tag{3.4}$$

where $S(\boldsymbol{\beta})$ can be expressed as

$$\begin{aligned} S(\boldsymbol{\beta}) &= \mathbf{y}^\top \mathbf{y} - \boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{y} - \mathbf{y}^\top \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{y}^\top \mathbf{y} - 2\boldsymbol{\beta}^\top \mathbf{X}\mathbf{y} + \boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta}, \end{aligned}$$

note that since $\boldsymbol{\beta}^\top \mathbf{X}\mathbf{y}$ is a scalar, its transpose is the same scalar. Taking the derivative respect to the $\boldsymbol{\beta}$ vector (Appendix **??** for details about matrix differentiation) and setting equal to zero

$$\frac{\partial S}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^\top \mathbf{y} + 2\mathbf{X}^\top \mathbf{X}\hat{\boldsymbol{\beta}} = 0,$$

which can be simplified to

$$\mathbf{X}^\top \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^\top \mathbf{y}, \tag{3.5}$$

the Equation (**??**) are the least squares normal equations. To solve this expression $\mathbf{X}^\top \mathbf{X}$ have to be invertible (this will be always true if the covariates in our model are linearly independent). We then have

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}. \tag{3.6}$$

It might happen that $\mathbf{X}$ is not full rank. This occur if, two of the explanatory variables are perfectly correlated (e.g., $x_{i2} = 3 \times x_{i4}$). This means that $\mathbf{X}\mathbf{X}^\top$ is singular and the least squares coefficients $\hat{\boldsymbol{\beta}}$ are not uniquely defined. In other words we are incorporating the same information twice. This problem it is usually resolved by recoding or dropping redundand columns in the design matrix $\mathbf{X}$ or applying variable selection methods. However, it is advisable to carry out a preliminary exploratory analysis in order to identify interactions between explanatory and possible correlations.

The fitted values can be obtained by

$$\hat{y}_i = \mathbf{x}_i^\top \hat{\boldsymbol{\beta}} = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_{ij},$$

The vector of fitted values can be obtained considering the entire design matrix

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} = \mathbf{H}\mathbf{y}, \tag{3.7}$$

We can notice that $\mathbf{H} \in \mathbb{R}^{n \times n}$ is usually called the hat matrix. It maps the vector of observed

(a) 3D Scatterplot

(b) Regression Hyperplane

Figure 3.2: A linear model fitted using least square. The observations are shown in red and the blue plane indicates the least squares fit to the data

values into a vector of fitted values.

The difference between the observed values y_i and the correnspondingfitted values $\hat{y}_i$ is the residual $e_i = y_i - \hat{y}_i$. The vector can be obtained

$$\mathbf{e} = \mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = \mathbf{y} - \hat{\mathbf{y}}. \tag{3.8}$$

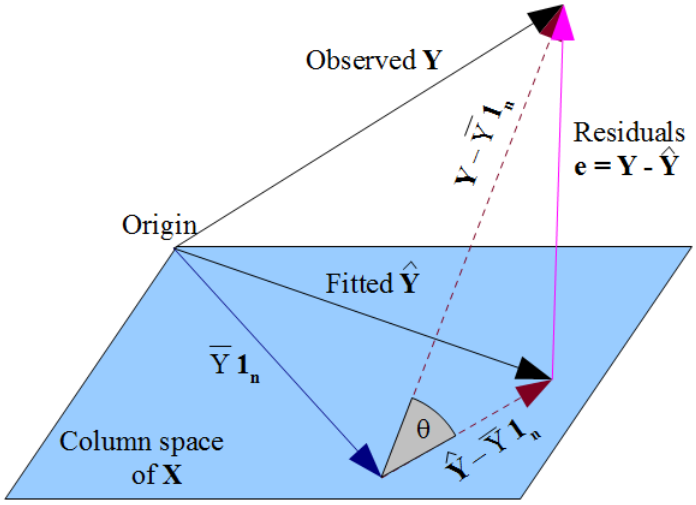


Figure 3.3: Geometric representation of linear regression. The figure illustrates the relationship between the observed outcomes Y, the fitted outcomes $\hat{Y}$, and the residuals e. The column space of X represents all possible fitted values, and the residuals are orthogonal to this space, indicating the difference between the observed and predicted outcomes. Source: Link

3.2 TODO(Optional) Gradient Descent Estimation

3.2.1 Properties

Denoting with $\mathbf{B}$ the estimator of $\boldsymbol{\beta}$, assuming that $\mathbb{E}(\boldsymbol{\epsilon}) = \mathbf{0}$, we can prove that $\mathbf{B}$ is an unbiased estimator of $\boldsymbol{\beta}$.

Proof.

$$\begin{aligned}
\mathbb{E}(\mathbf{B}) &= \mathbb{E}[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}] \\
&= \mathbb{E}[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})] \\
&= \mathbb{E}[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\epsilon}] \\
&= \boldsymbol{\beta} + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbb{E}[\boldsymbol{\epsilon}] \\
&= \boldsymbol{\beta},
\end{aligned}$$

□

we notice that when $\mathbb{E}(\boldsymbol{\epsilon}) \neq \mathbf{0}$ our estimator is bound to be biased and the amount of $bias = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbb{E}[\boldsymbol{\epsilon}]$ will depend on the design matrix and by the expected value of the error terms.

The covariance matrix of $\mathbf{B}$, assuming that the error terms are uncorrelated and that the $Var(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$, is

$$Var(\mathbf{B}) = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}.$$

Proof.

$$\begin{aligned}
Var(\mathbf{B}) &= Var[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}] \\
&= Var[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})] \\
&= Var[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\epsilon}] \\
&= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top Var(\boldsymbol{\epsilon})\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}.
\end{aligned}$$

□

To obtain the sampling distribution of $\mathbf{B}$ we need to assume that $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$. *So far the expressions of our estimators has been obtained as just an albegraic tool, no Normality assumption is required to obtain* $\mathbf{B}$. This simply means that properties A1-A3 are only useful to leverage the properties of estimator $\mathbf{B}$. Assuming that the erros are normally distributed we can obtain that

$$\mathbf{B} \sim N_p(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}),$$

where N_p denotes a $p-$variate normal distribution. The fact of assuming that $\boldsymbol{\epsilon}$ are normally distributed ensure us to have that $\mathbf{B}$ is normal as well. This result comes from noting that $\mathbf{B}$ is a linear transformation of $\mathbf{Y}$. In fact

$$\boldsymbol{B} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}).$$

Observation Number	Delivery Time, $y(\min)$	Number of Cases, x_1	Distance, $x_2(\text{ft})$
1	16.68	7	560
2	11.50	3	220
3	12.03	3	340
4	14.88	4	80
5	13.75	6	150
6	18.11	7	330
7	8.00	2	110
8	17.83	7	210
9	79.24	30	1460
10	21.50	5	605
11	40.33	16	688
12	21.00	10	215
13	13.50	4	255
14	19.75	6	462
15	24.00	9	448
16	29.00	10	776
17	15.35	6	200
18	19.00	7	132
19	9.50	3	36
20	35.10	17	770
21	17.90	10	140
22	52.32	26	810
23	18.75	9	450
24	19.83	8	635
25	10.75	4	150

Table 3.1: Delivery Time Data.

Example 5. *A soft drink bottler is analyzing the vending machine service routes in his distribution system. He is interested in predicting the amount of time required by the route driver to service the vending machines in an outlet. This service activity includes stocking the machine with beverage products and minor maintenance or housekeeping. The industrial engineer responsible for the study has suggested that the two most important variables affecting the delivery time* (y) *are the number of cases of product stocked* (x_1) *and the distance walked by the route driver* (x_2)*. The engineer has collected 25 observations on delivery time, which are shown in Table* **??***. We will fit the multiple linear regression model*

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon,$$

to the delivery time data in Table **??***. When there are only two regressors, sometimes a three-dimensional scatter diagram is useful in visualizing the relationship between the response and the regressors. By spinning these plots, some software packages permit different views of the point*

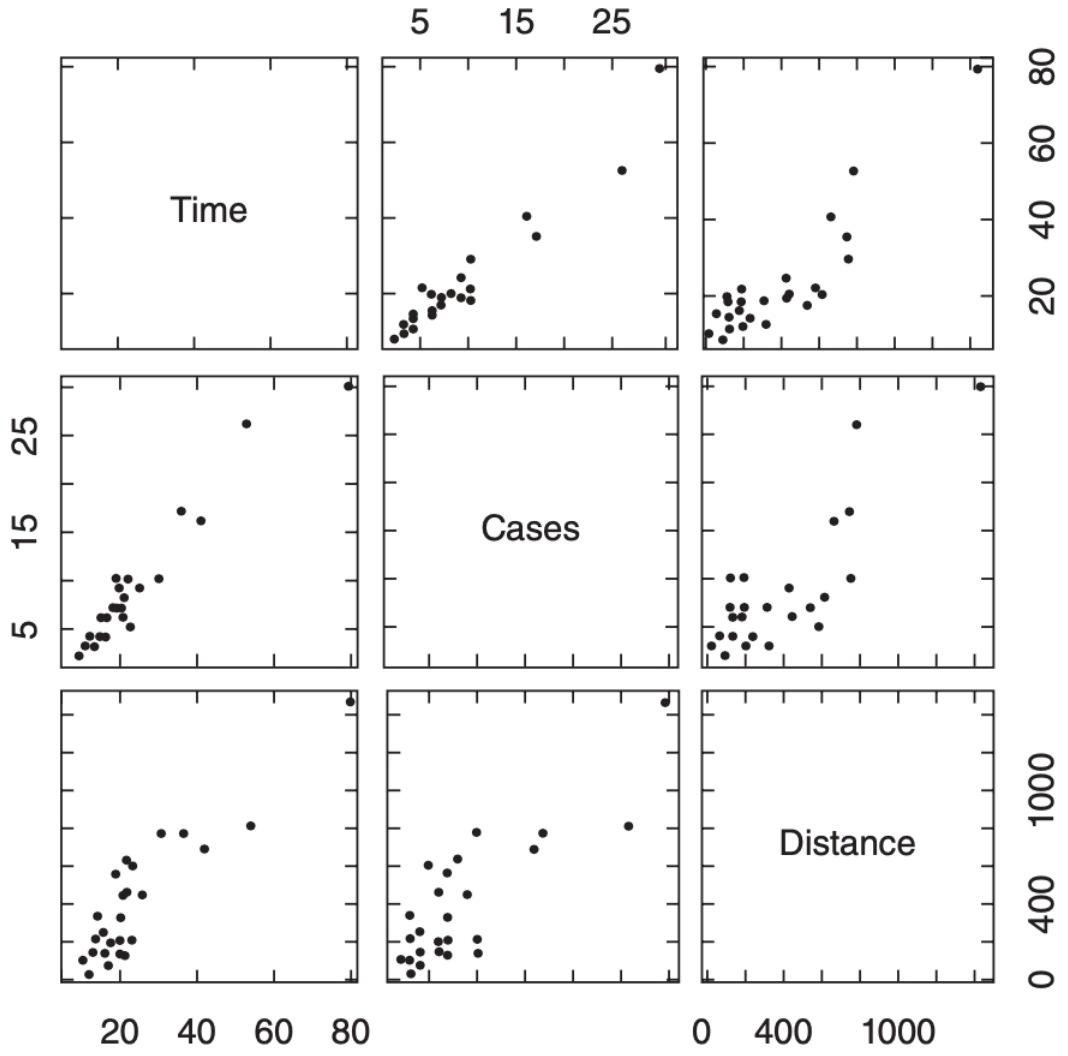


Figure 3.4: Example 5 data scatterplot. We observe how there seems to be a linear relationship between time and the two covariates (cases and distance). It seems reasonable to estimate a linear model.

cloud. This view provides an indication that a multiple linear regression model may provide a reasonable fit to the data.

To fit the multiple regression model we first form the $\mathbf{X}$ *matrix and* $\mathbf{y}$ *vector:*

$$\mathbf{X} = \begin{bmatrix} 1 & 7 & 560 \\ 1 & 3 & 220 \\ 1 & 3 & 340 \\ 1 & 4 & 80 \\ 1 & 6 & 150 \\ 1 & 7 & 330 \\ 1 & 2 & 110 \\ 1 & 7 & 210 \\ 1 & 30 & 1460 \\ 1 & 5 & 605 \\ 1 & 16 & 688 \\ 1 & 10 & 215 \\ 1 & 4 & 255 \\ 1 & 6 & 462 \\ 1 & 9 & 448 \\ 1 & 10 & 776 \\ 1 & 6 & 200 \\ 1 & 7 & 132 \\ 1 & 3 & 36 \\ 1 & 17 & 770 \\ 1 & 10 & 140 \\ 1 & 26 & 810 \\ 1 & 9 & 450 \\ 1 & 8 & 635 \\ 1 & 4 & 150 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} 16.68 \\ 11.50 \\ 12.03 \\ 14.88 \\ 13.75 \\ 18.11 \\ 8.00 \\ 17.83 \\ 79.24 \\ 21.50 \\ 40.33 \\ 21.00 \\ 13.50 \\ 19.75 \\ 24.00 \\ 29.00 \\ 15.35 \\ 19.00 \\ 9.50 \\ 35.10 \\ 17.90 \\ 52.32 \\ 18.75 \\ 19.83 \\ 10.75 \end{bmatrix}$$

The $\mathbf{X}^\top\mathbf{X}$ *matrix is*

$$\mathbf{X}^\top\mathbf{X} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 7 & 3 & \cdots & 4 \\ 560 & 220 & \cdots & 150 \end{bmatrix} \begin{bmatrix} 1 & 7 & 560 \\ 1 & 3 & 220 \\ \vdots & \vdots & \vdots \\ 1 & 4 & 150 \end{bmatrix} = \begin{bmatrix} 25 & 219 & 10,232 \\ 219 & 3,055 & 133,899 \\ 10,232 & 133,899 & 6,725,688 \end{bmatrix}$$

and the $\mathbf{X}^\top\mathbf{y}$ *vector is*

$$\mathbf{X}^\top \mathbf{y} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 7 & 3 & \cdots & 4 \\ 560 & 220 & \cdots & 150 \end{bmatrix} \begin{bmatrix} 16.68 \\ 11.50 \\ \vdots \\ 10.75 \end{bmatrix} = \begin{bmatrix} 559.60 \\ 7,375.44 \\ 337,072.00 \end{bmatrix}$$

The least-squares estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}$$

or

$$\begin{aligned} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} &= \begin{bmatrix} 25 & 219 & 10,232 \\ 219 & 3,055 & 133,899 \\ 10,232 & 133,899 & 6,725,688 \end{bmatrix}^{-1} \begin{bmatrix} 559.60 \\ 7,375.44 \\ 337,072.00 \end{bmatrix} \\ &= \begin{bmatrix} 0.11321518 & -0.00444859 & -0.00008367 \\ -0.00444859 & 0.00274378 & -0.00004786 \\ -0.00008367 & -0.00004786 & 0.00000123 \end{bmatrix} \begin{bmatrix} 559.60 \\ 7,375.44 \\ 337,072.00 \end{bmatrix} \\ &= \begin{bmatrix} 2.34123115 \\ 1.61590712 \\ 0.01438483 \end{bmatrix} \end{aligned}$$

The least-squares fit (with the regression coefficients reported to five decimals) is

$$\hat{y} = 2.34123 + 1.61591x_1 + 0.01438x_2.$$

3.2.2 Gauss Markov Theorem

We can now discuss one of the result that we are going to see in this course. **Gauss- Markov theorem**.

Theorem 3. *(Gauss- Markov Theorem). If the assumptions of the linear regression model hold, then the ordinary least squares (OLS) estimator* $\mathbf{B}$ *is the Best Linear Unbiased Estimator (BLUE) of* $\boldsymbol{\beta}$*. This means that among all linear and unbiased estimators of the coefficients, the OLS estimator has the smallest variance. Hence the most efficient estimator.*

Proof. First of all we can notice that

$$\begin{aligned}\mathbf{B} &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}\\ &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon})\\ &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}+(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\epsilon}\\ &= \boldsymbol{\beta}+\mathbf{A}\boldsymbol{\epsilon},\end{aligned}$$

where $\mathbf{A} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ such that $\mathbf{A}\mathbf{A}^\top = (\mathbf{X}^\top\mathbf{X})^{-1}$, the OLS estimators can be written as

$$\mathbf{B} = \mathbf{A}\mathbf{Y} = \boldsymbol{\beta}+\mathbf{A}\boldsymbol{\epsilon}.$$

The estimator $\mathbf{B}$ for the parameters $\boldsymbol{\beta}$ are linear as $\mathbf{B} = \mathbf{A}\mathbf{Y}$ is a linear combination of the random variables $\mathbf{Y}$. The variance can be expressed as

$$Var(\mathbf{B}) = Var(\mathbf{A}\mathbf{Y}) = \mathbf{A}\,\mathrm{Var}(\mathbf{Y})\mathbf{A}^\top = \sigma^2\mathbf{A}\mathbf{A}^\top = \sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1},$$

now we need to show that OLS estimators are the most efficient in the class of the linear unbiased estimators. We introduce, for a given matrix $\mathbf{C}$, e generic linear and unbiased estimator for $\boldsymbol{\beta}$

$$\tilde{\mathbf{B}} = (\mathbf{A}+\mathbf{C})\mathbf{Y} = \mathbf{A}\mathbf{Y}+\mathbf{C}(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}) = \mathbf{B}+\mathbf{C}\mathbf{X}\boldsymbol{\beta}+\mathbf{C}\boldsymbol{\epsilon}.$$

As we are assuming that $\tilde{\mathbf{B}}$ is unbiased, $\mathbb{E}(\tilde{\mathbf{B}}) = \boldsymbol{\beta}$, this is equivalent to imposing

$$\mathbb{E}(\tilde{\mathbf{B}}) = \mathbb{E}(\mathbf{B})+\mathbf{C}\mathbf{X}\boldsymbol{\beta}+\mathbf{C}\mathbb{E}(\boldsymbol{\epsilon}) = \boldsymbol{\beta}+\mathbf{C}\mathbf{X}\boldsymbol{\beta} = [\mathbf{I}+\mathbf{C}\mathbf{X}]\boldsymbol{\beta} = \boldsymbol{\beta} \Leftrightarrow \mathbf{C}\mathbf{X} = \mathbf{O}.$$

The constraint $\mathbf{C}\mathbf{X} = \mathbf{O}$ implies:

$$\mathbf{C}\mathbf{A}^\top = \mathbf{C}\left[\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\right]^\top = \mathbf{C}\left[\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\right] = \mathbf{C}\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1} = \mathbf{O}\left(\mathbf{X}'\mathbf{X}\right)^{-1} = \mathbf{O},$$

therefore

$$\mathbf{O} = \mathbf{C}\mathbf{A}^\top = \left(\mathbf{A}\mathbf{C}^\top\right)^\top \Rightarrow \mathbf{A}\mathbf{C}^\top = \mathbf{O}.$$

The variance and covariance matrix of $\tilde{\mathbf{B}}$ is:

$$\begin{aligned} Var(\tilde{\mathbf{B}}) &= Var[(\mathbf{A}+\mathbf{C})\mathbf{Y}] = (\mathbf{A}+\mathbf{C})Var(\mathbf{Y})(\mathbf{A}+\mathbf{C})^\top \\ &= (\mathbf{A}+\mathbf{C})\sigma^2\mathbf{I}(\mathbf{A}+\mathbf{C})^\top = \\ &= \sigma^2\left(\mathbf{A}\mathbf{A}^\top + \mathbf{C}\mathbf{A}^\top + \mathbf{A}\mathbf{C}^\top + \mathbf{C}\mathbf{C}^\top\right) \\ &= \sigma^2\left(\mathbf{A}\mathbf{A}^\top + \mathbf{C}\mathbf{C}^\top\right) = \sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1} + \sigma^2\left(\mathbf{C}\mathbf{C}^\top\right) = \\ &= Var(\mathbf{B}) + \sigma^2\left(\mathbf{C}\mathbf{C}^\top\right). \end{aligned}$$

Therefore, $Var(\tilde{\mathbf{B}}) - Var(\mathbf{B}) = \sigma^2\left(\mathbf{C}\mathbf{C}^\top\right)$ is a positive semi-definite matrix. We can conclude that, the variance of the OLS estimators of any regression parameter will not exceed the variance of any other linear and unbiased estimator for $\boldsymbol{\beta}$. The two variances will coincide if and only if $\mathbf{C} \equiv \mathbf{O}$, i.e., when: $\tilde{\mathbf{B}} \equiv \mathbf{B}$.

Hence, we have shown that the OLS estimators are BLUE.

□

3.2.3 Maximum Likelihood Estimation

The least square estimator is also a maximum likelihood estimator (MLE) when the rensponses are independent and identically distributed. When $Y_i \sim N(\mathbf{x}_i^\top \boldsymbol{\beta}, \sigma^2)$, therefore the density function is

$$f_{Y_i}(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(- \frac{(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^\top (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})}{2\sigma^2} \right),$$

once have observed a realization of $(Y_1, \ldots, Y_n)$, the likelihood function is

$$L(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}) = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp\left(- \frac{(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2} \right),$$

where the maximization of $L(\boldsymbol{\beta}, \sigma^2 | \mathbf{y})$ with respect to $\boldsymbol{\beta}$ for given σ^2 is equivalent to maximize the exponential part whose is maximum when $(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$ is miminum. *Please attempt to obtain the MLE for $\boldsymbol{\beta}$ and for σ^2 as exercise.*

3.2.4 Residuals

In assessing the fit of out model we need to consider the behaviour of the residuals. With fitted values we denote the estimates of the mean rensponse for each observation $\hat{y}_i = \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}$ for $i = 1, \ldots, n$. So the residuals can be defined as

$$e_i = y_i - \hat{y}_i = y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}},$$

in matrix notation

$$\begin{aligned} \mathbf{e} &= \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y} - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \\ &= \mathbf{y} - \mathbf{H}\mathbf{y} \\ &= (\mathbf{I}_n - \mathbf{H})\mathbf{y}, \end{aligned}$$

where $\mathbf{I}_n$ is a $n \times n$ identity matrix and $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$ is the hat matrix. $\mathbf{H}$ is called in this way as it puts the hat on $\mathbf{y}$. In fact if we multiply $\mathbf{y}$ by $\mathbf{H}$ we obtain $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$.

We can denote $\mathbf{e}^\top = (e_1, \ldots, e_n)$ the vector containing the r.v. residuals generated by $\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$. Let $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$, $\mathbf{M} = \mathbf{I} - \mathbf{H} = \mathbf{I} - \mathbf{X}\mathbf{A}$ and $\mathbf{A} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$ then

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y} - \mathbf{X}(\mathbf{A}\mathbf{y}) = (\mathbf{I} - \mathbf{H})\mathbf{y} = \mathbf{M}\mathbf{y},$$

$\mathbf{M}$ is an idempotent matrix with the following properties

$$\mathbf{M} = \mathbf{M}^\top; \quad \mathbf{M}\mathbf{M} = \mathbf{M}; \quad \mathbf{M}^\top \mathbf{M} = \mathbf{M}; \quad \text{rank}(\mathbf{M}) = \text{tr}(\mathbf{M}) = n - p - 1.$$

Furthermore $\mathbf{MX} = \mathbf{X}^\top\mathbf{M} = \mathbf{0}$

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{My} = \mathbf{M}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = \mathbf{0}\boldsymbol{\beta} + \mathbf{M}\boldsymbol{\epsilon}.$$

The residual random variable is a linear transformation of an error terms.

$$\mathbb{E}(\mathbf{e}) = \mathbb{E}(\mathbf{M}\boldsymbol{\epsilon}) = \mathbf{M}\mathbb{E}(\boldsymbol{\epsilon}),$$

and

$$\begin{aligned}
\mathbb{E}(\mathbf{e}^\top\mathbf{e}) &= \mathbb{E}[(\mathbf{M}\boldsymbol{\epsilon})^\top(\mathbf{M}\boldsymbol{\epsilon})] \\
&= \mathbb{E}[\boldsymbol{\epsilon}^\top\mathbf{M}^\top\mathbf{M}\boldsymbol{\epsilon}] \\
&= \mathbb{E}[\text{tr}(\boldsymbol{\epsilon}^\top\mathbf{M}^\top\mathbf{M}\boldsymbol{\epsilon})] \\
&= \mathbb{E}[\text{tr}(\mathbf{M}^\top\mathbf{M}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top)] \\
&= \text{tr}(\mathbb{E}[\mathbf{M}^\top\mathbf{M}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top]) \\
&= \text{tr}(\mathbf{M}^\top\mathbf{M}\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top]) \\
&= \sigma^2\text{tr}(\mathbf{M}^\top\mathbf{M}) \\
&= \sigma^2\text{tr}(\mathbf{M}) \\
&= \sigma^2(n-p-1).
\end{aligned}$$

3.2.5 Inference on σ^2

Once $\boldsymbol{\beta}$ have been estimated, we can obtain the residual sum of squares (RSS), then

$$RSS = \mathbf{e}^\top\mathbf{e} = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}),$$

we have proved that

$$\mathbb{E}(RSS) = (n-p-1)\sigma^2$$

Hence an unbiased estimator of σ^2 result to be

$$\hat{\sigma}^2 = \frac{RSS}{(n-p-1)}$$

this since $\mathbb{E}(\hat{\sigma}^2) = \mathbb{E}(RSS)/(n-p-1) = (n-p-1)\sigma^2/(n-p-1) = \sigma^2$. Assuming that $\epsilon_i \sim N(0, \sigma^2)$, it can be shown that $\frac{RSS}{\sigma^2} \sim \chi^2_{n-p}$. Where with χ^2 we denoted a chi-squared random variable. Click here for more details here.

If we assume that the random variable $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ is distributed normally, then the residuals

will be distributed as

$$\mathbf{e} \sim N(\mathbf{0}, \sigma^2(\mathbf{I} - \mathbf{H})).$$

3.2.6 Interpetation

In multiple linear regression, the coefficients represent the estimated effect of each predictor variable (x_i's) on the outcome variable (y_i), while holding all other predictors constant. Specifically, the intercept (β_0) represents the estimated mean value of the outcome variable when all predictor variables are equal to zero. In many cases, this intercept may not have a meaningful interpretation, particularly if the predictor variables do not take on meaningful zero values.

The coefficients for the predictor variables ($\beta_1, \beta_2, \ldots, \beta_p$) represent the estimated change in the outcome variable for a one-unit increase in each predictor variable, holding all other predictors constant. These coefficients can be interpreted as the "effect size" of each predictor variable on the outcome variable. For example, if we have a multiple linear regression model that predicts the salary of employees based on their years of experience and level of education, we can interpret the coefficient for years of experience as the estimated increase in salary associated with a one-year increase in experience, holding education constant. Similarly, we can interpret the coefficient for education as the estimated increase in salary associated with a one-level increase in education, holding experience constant.

We can assume to have the following linear model

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i \;\; i = 1 \ldots, n,$$

where $\mathbf{x}_i^\top = (1, x_{i1}, \ldots, x_{in})^\top$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k, \ldots, \beta_p)^\top$, in this case β_k measures the expected change in y_i if x_{ik} changes with one unit, whereas the other variables in $\mathbf{x}_i$ do not change. Formally

$$\frac{\partial \mathbb{E}(y_i)}{\partial x_{ik}} = \beta_k.$$

It is important to realize that we had to state explicitly that the other variables in x_i did not change. In a multiple regression model, single coefficients can only be interpreted under ceteris paribus conditions. If $\mathbf{x}_i^\top \boldsymbol{\beta}$ contains , say $x_i\beta_2 + x_i^2\beta_3$. With abuse of notation, taking the derivative we have

$$\frac{\partial \mathbb{E}(y_i)}{\partial x_i} = \beta_2 + 2x_i\beta_3,$$

which can be interpreted as the marginal effect of a changing x_i if the other variables (excluding x_i^2) are kept constant. This shows how the marginal effects of explanatory variables can be allowed to vary over the observations by including additional terms involving these variables (in

this case x_i^2).

In the case of transformations, interpretations change slightly. Here are some examples:

Model	**Interpretation**
$y_i = \beta_0 + \beta_1 \log(x_{i1}) + \epsilon_i$	A 1% change in x is associated with an expected change in y of $0.01 \times \beta_1$.
$\log(y_i) = \beta_0 + \beta_1 x_{i1} + \epsilon_i$	A change in x by one unit ($\nabla x = 1$) is associated with an expected change in y of $(\exp(\beta_1) - 1) \times 100$.
$\log(y_i) = \beta_0 + \beta_1 \log(x_{i1}) + \epsilon_i$	A change of 1% in x is associated with a $\beta_1\%$ change in y.

Table 3.2: Models and Their Interpretations.

3.3 The Geometry of Transformations

Response transformation introduce Normality of a distribution and stabilize variances because they can stretch apart data in one region and push observations together in other regions. Plot **??** illustrates this behavior. On the horizontal axis is a sample of data from a right skewed logNormal distribution. The transformation $h(y)$ is the logarithm. The symmetry is achieved because the log transformation stretches apart data with small values and shrinks together data with large values. This becomes evident observing the derivative of the log funtion. The derivative of $log(y)$ is $1/y$ which is a decreasing function of y.

Consides an arbitrary increasing function, $h(y)$. If x and x' are two nearby data points that are transformed to $h(x)$ and $h(x')$ respectively then the distance between transformed values can be obtained by expanding $h(x)$ around x'

$$h(x) \approx h(x') + h'(x - x')$$

adjusting the terms and taking the absolute values

$$|h(x) - h(x')| \approx h'|x - x'|$$

therefore $h(x)$ and $h(x')$ are stretched apart where h' is large. $h(x)$ and $h(x')$ are stretched apart when h' is small. A function $h(\cdot)$ is called concave if $h'(x)$ is a decreasing function of x.

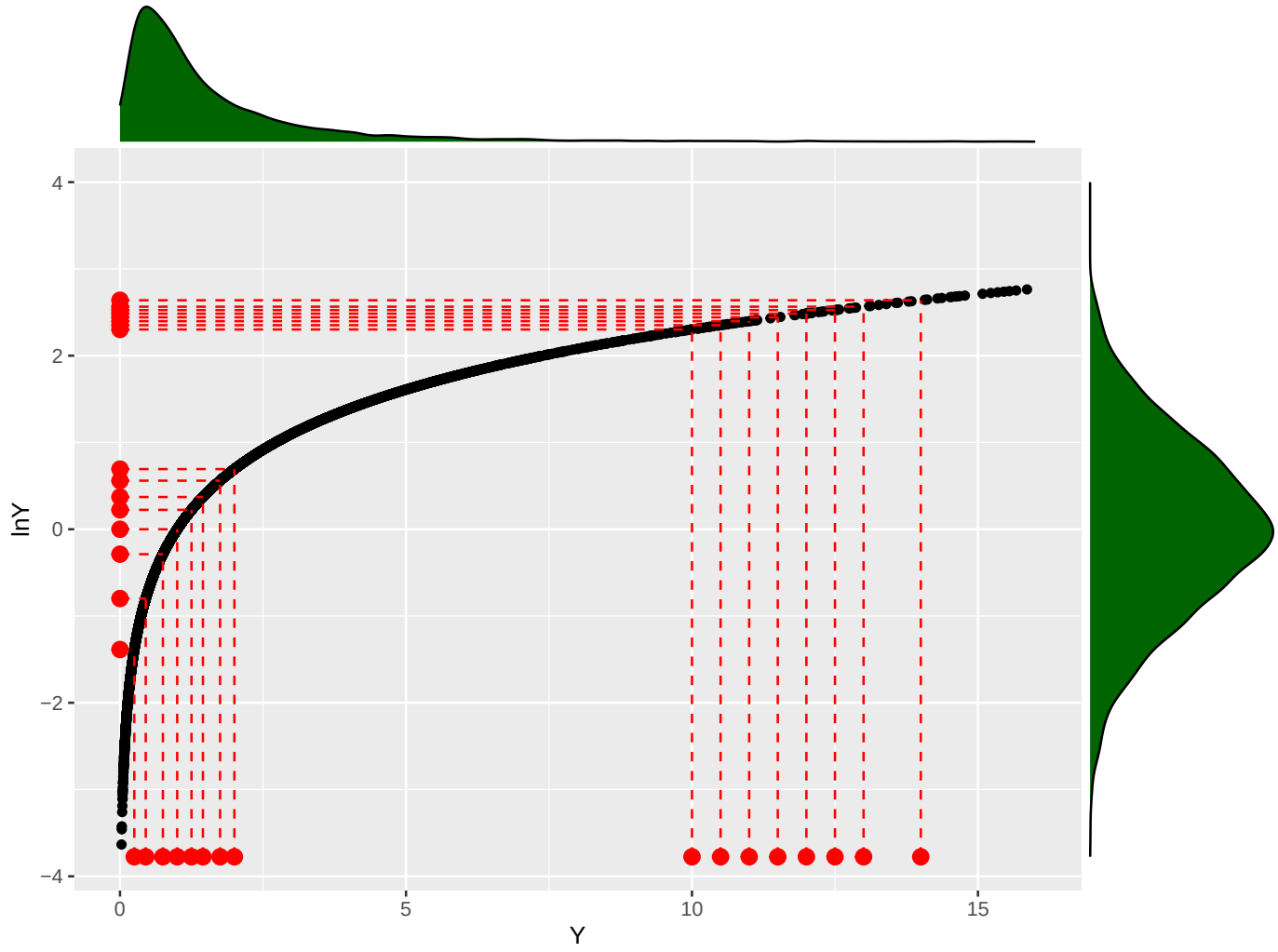


Figure 3.5: A symmetrizing transformation. The skewed lognormal data on the horizontal axis are transformed to symmetry by the log transformation.

3.4 Inference

3.4.1 Confidence Intervals

In case we use the assumption of $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$ (this assumption should be checked carefully using the residuals), if we assume that $\hat{\beta}_j$ is the j^{th} element of $\hat{\boldsymbol{\beta}}$ we can derive that

$$\hat{\beta}_j \sim \mathcal{N}(\beta_j, \sigma^2 v_j),$$

where v_j is the $(j,j)^{th}$ diagonal element of $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$. Recalling that

$$RSS/\sigma^2 \sim \chi^2_{n-(p+1)},$$

we have

$$T = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 v_j}} \sim t_{n-(p+1)},$$

where $\hat{\sigma} = RSS/(n-(p+1))$

An exact $100(1-\alpha)\%$ confidence interval for β_j is

$$\hat{\beta}_j \pm t_{n-(p+1),\frac{1}{2}\alpha} se(\hat{\beta}_j),$$

where $se(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 v_j}$ and $t_{n-(p+1),\frac{1}{2}\alpha}$ is the upper $100 \times \frac{1}{2}\alpha\%$ quantile of the $t_{n-(p+1)}$ distribution. This result can be used to test hypotheses of the form $H_0 : \beta_j = b_0$ for a given j. In particular when $H_0 : \beta_j = 0$. The test statistics under H_0 reduces to

$$\frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \sim t_{n-(p+1)},$$

from which we can obtain p-values. Here we are testing whether the response depends on the associated explanatory variable, given the inclusion of the other explanatory variables in the model.

3.4.2 Tests

3.4.3 Testing one linear restriction

Often, it is of interest to test more than one coefficient, such as $\beta_1 + \beta_2 + \ldots, \beta_p = 1$ in general such linear hypothesis can be formulated, without loss of generality, in the following way

$$H_0 : r_1\beta_1 + r_2\beta_2 + \cdots + r_p\beta_p = \mathbf{r}^\top \boldsymbol{\beta} = q$$

where $q \in \mathbb{R}$ (is a scalar value), $\mathbf{r} \in \mathbb{R}^p$. Using the result of the Gauss Markov theorem

and the fact that also $\mathbf{r}^\top\mathbf{B}$ is the Best Linear Unbiased Estimator (BLUE) for $\mathbf{r}^\top\boldsymbol{\beta}$ with variance $V(\mathbf{r}^\top\mathbf{B}) = \mathbf{r}^\top Var(\mathbf{B})\mathbf{r}$ and then $se(\mathbf{r}^\top\mathbf{B}) = \sqrt{V(\mathbf{r}^\top\mathbf{B})}$. Replacing σ^2 with its estimate and noting that $\mathbf{B}$ is a p variate normal distribution, $\mathbf{r}^\top\mathbf{B}$ is normal as well. Therefore

$$T = \frac{\mathbf{r}^\top\mathbf{B} - \mathbf{r}^\top\boldsymbol{\beta}}{se(\mathbf{r}^\top\mathbf{B})} \sim t_{n-p}$$

Under the null we have

$$T|H_0 \text{ is true } = \frac{\mathbf{r}^\top\mathbf{B} - q}{se(\mathbf{r}^\top\mathbf{B})}$$

this test statistic has a distribution t-student with $n - p$ degree of freedom (d.o.f).

3.4.4 Test about multiple parameters

If we consider the null that a subset of $(p + 1) - q$ parameters out of $(p + 1)$ are zero. Let H_1 denote the alternative hypothesis that all p parameters are not 0 and let RSS_0 and RSS_1 denote the residual sum of squares under H_0 and H_1, respectively. Under the null

$$RSS_0/\sigma^2 \sim \chi^2_{n-(p+1)}$$

this result combined with the fact that $RSS_0 - RSS_1$ and RSS are independent. Yields that under the null

$$\frac{RSS_0 - RSS_1}{\sigma^2} \sim \chi^2_{(p+1)-q}$$

we obtain the F-test

$$\frac{RSS_0 - RSS_1/(p + 1 - q)}{RSS_1/(n - p - 1)}$$

under the null follows a Fisher distribution with $p + 1 - q$ and $n - p - 1$ d.o.f.

3.4.5 Testing: further topics

We an use the results of the previous sections to formulate an even more general test. The most general linear null hypothesis is a combination of the previous two cases and comprises a set of J linear restrictions on the coefficients, this can be formulated as

$$\mathbf{R}\boldsymbol{\beta} = \mathbf{q}$$

where $\mathbf{R} \in \mathbb{R}^{J\times p+1}$, assumed to be **full rank** , and $\mathbf{q} \in \mathbb{R}^J$. If for example we want to test

$\beta_1 + \beta_2 + \cdots + \beta_p = 1$ and $\beta_2 = \beta_3$, $\mathbf{R}$ and $\mathbf{q}$ have to be specified in the following way

$$\mathbf{R} = \begin{pmatrix} 0 & 1 & 1 & \dots & \dots & 1 \\ 0 & 1 & -1 & 0 & \dots & 0 \end{pmatrix}, \quad \mathbf{q} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

Using the assumption that $\boldsymbol{\epsilon}$ is normally distributed, we conclude that $\mathbf{RB}$ is normally distributed with mean $\mathbf{R}\boldsymbol{\beta}$ and variance $\mathbf{R}V(\mathbf{B})\mathbf{R}^\top$. This means that, under the null, the quadratic form

$$(\mathbf{RB} - \mathbf{q})^\top V(\mathbf{RB})^{-1}(\mathbf{RB} - \mathbf{q})$$

has a Chi-squared distribution with J degrees of freedom. Unfortunatelly, σ^2 is unknown so we need to substitute it with an unbiased estimate $\hat{\sigma}^2$. The resulting test statistic is

$$\psi = (\mathbf{RB} - \mathbf{q})^\top [\mathbf{R}\hat{V}(\mathbf{B})\mathbf{R}^\top]^{-1}(\mathbf{RB} - \mathbf{q}) \tag{3.9}$$

where $\hat{V}(\mathbf{B}) = \hat{\sigma}^2(\mathbf{X}^\top\mathbf{X})^{-1}$. In large samples, the difference between σ^2 and it estimates has limited impact and the test statistic in **??** is approximately distributed as a Chi-squared distribution. It can be proved that an exact distribution under the the usual assumptions we can derive an exact statistics

$$F = \frac{(\mathbf{RB} - \mathbf{q})^\top [\mathbf{R}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}^\top]^{-1}(\mathbf{RB} - \mathbf{q})}{J\hat{\sigma}^2}$$

under the null F follows a Fisher distribution with J and $n - (p+1)$ d.o.f. As before, large values of F lead to rejection of the null (can you see why?).

3.4.6 Prediction

So far we have seen how to estimate a linear model and how to make inference. The model that we just estimated can be used to predict the future... lets assume to have observed a new vector of covariates $\mathbf{x}_0 = (\mathbf{x}_{01}, \mathbf{x}_{02}, \ldots, \mathbf{x}_{0p})^\top$ then a **point estimate of the future observation** y_0 at the point $\mathbf{x}_0$ is

$$\hat{y}_0 = \mathbf{x}_0^\top \hat{\beta}$$

where $\hat{\boldsymbol{\beta}}$ has been estimated not including the new information $\mathbf{x}_0$, therefore $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$. However, in this case a point estimate is not sufficient, we need also to asses the uncertainty in this prediction. A $100(1-\alpha)\%$ prediction interval for this future observation is

$$\hat{y}_0 \pm t_{n-(p+1)\frac{1}{2}\alpha}\sqrt{\hat{\sigma}^2\left(1 + \mathbf{x}_0^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_0\right)}.$$

We can also be interested in obtaining a **prediction of the mean response**. In this case the $100(1-\alpha)\%$ is

$$\hat{y}_0 \pm t_{n-(p+1)\frac{1}{2}\alpha}\sqrt{\hat{\sigma}^2\mathbf{x}_0^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_0},$$

further application details will be discussed in the labs.

3.5 Regression with qualitative variables

The linear regression theory that we have studied so far can be generalized to the case of quantitative variables. Sometimes, we may need to employ quantitative variables in the analysis, such as operators, employment status (employed or unemployed), shifts (day, evening, or night), and sex (male or female). In general, a qualitative variable does not have a natural scale of measurement. We assign a set of levels to a qualitative variable to account for the effect that the variable may have on the response. This is done through the use of indicator variables. In econometrics, indicator variables are sometimes referred to as dummy variables. Let's assume we have the following dummy variable.

$$\mathbf{1}_{\{Female\}} = \begin{cases} 0 & \text{if the observation is Male} \\ 1 & \text{if the observation is Female} \end{cases}$$

$\mathbf{1}_{\{Female\}}$ is a random variable such that assumes value 1 when the ith observation is female and 0 otherwise. We used 0,1 is an arbitrary way. We can assume to have the following model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 \mathbf{1}_{\{Female\}} + \varepsilon. \tag{3.10}$$

To interpret this model we need to consider the case in which the observations is a male. Therefore,

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2(0) + \varepsilon \\ &= \beta_0 + \beta_1 x_1 + \varepsilon. \end{aligned} \tag{3.11}$$

Thus, the relationship between y and x_1 when the observation is male turn out to be a straight line with intercept β_0 and slope β_1. For female we have

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2(1) + \varepsilon \\ &= (\beta_0 + \beta_2) + \beta_1 x_1 + \varepsilon. \end{aligned} \tag{3.12}$$

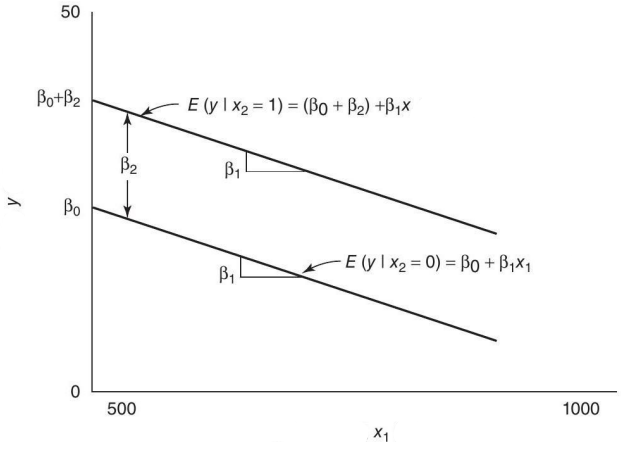


Figure 3.6: Rensponse function for different values of $\mathbf{1}_{\{Female\}}$.

That is, for females, the relationship between y and x_1 is also a straight line with a slope of β_1 but an intercept of $\beta_0 + \beta_2$.

The two response functions are shown in Figure **??**. Models (**??**) and (**??**) describe two parallel regression lines, which means they have a common slope of β_1 but different intercepts. Additionally, the variance of the errors ε is assumed to be the same for both female and male observations. The parameter β_2 represents the difference in heights between the two regression lines, indicating the measure of the difference in y when transitioning from male to female.

3.6 Checking Model Adequacy

Here are the crucial model assumptions of the multiple linear regression model:

(i) **Linearity of the relationship between predictors and Y.**

(ii) **Normality of errors** e_i. Though the normal distribution theoretically can generate arbitrary real numbers (including very large ones), very extreme values occur under the normal distribution with a very small probability. For example, the probability that a point is generated which is further than 5σ away from the mean value (the residuals are centered around 0) is about 5×10^{-7} which means that only one such point can be expected in more than 1.5 millions of points generated from the same normal distribution. Because outliers have a very strong influence on least squares regression, the detection of outliers is the most important task connected to the normality assumption about the ϵ_i. Another possible important deviation from normality could be skewness of the distributional shape (many points one one side but much fewer points, some of them appearing somewhat outlying, on the other side).

Some other deviations from normality are less dangerous. For example, under uniformly distributed random variation with restricted value range (this may hold, for example, if the y-variable consists of percentages), the normal theory still works quite well. Generally, you should have in mind that normality (as well as all the other model assumptions) is an idealization that never holds precisely in practice. The important question is not whether a distribution is really normal but whether there are deviations from normality that may lead the applied statistical methodology (here linear regression) to misleading conclusions about the data.

(iii) **Homogeneity of variances**("homoscedasticity" as opposed to "heteroscedasticity") of e_i. This implies particularly that the variances do not depend on any of the predictor variables.

(iv) **Independence of errors** e_i (of each other).

It is important to note that these assumptions are related to the residuals. Hence, we are not able to detect departures from the underlying assumptions by examining the standard summary statistics (e.g., R^2, RSS, t and F statistics). Since the residuals play a significant role, it is natural to assume that most of the diagnostic techniques will be based on examining residuals. Some useful plots are:

(i-a) **Matrix plot**: The so-called matrix plot (command in `R` is `pairs()`) consists of all scatterplots of any pair of predictor variables and predictors vs. response, arranged in matrix form. The matrix plot can (and should) be plotted without having fitted a linear regression (or any other model) before. The plots of predictor variables vs. the response can be used to assess linearity, outliers and homoscedasticity. Note, however, that there is some danger of over-interpreting impressions from these plots, because to see the response plotted against every single predictor variable does not give a full impression. For example, it is possible that an apparently nonlinear shape of the plot of a single predictor vs. the response is rather caused by values of the other predictors than by a real violation of linearity (though strong nonlinear or heteroscedastic shapes

hint at real violations of model assumptions in practice in most cases).

The plots of pairs of predictors can reveal collinearity and leverage points (see below), which are not violations of the model assumptions (there are no model assumptions about the predictor variables in linear regression), but still problematic.

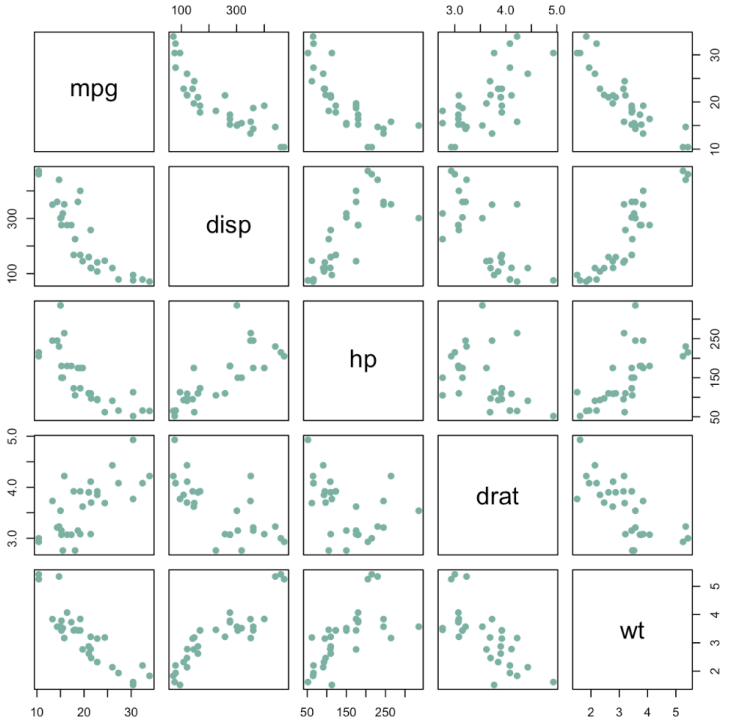


Figure 3.7: This is an example of a scatterplot matrix based on the `mtcars` dataset. Where `mpg`= Miles per gallon (fuel efficiency), `disp`= Displacement (cubic inches), `hp`= Horsepower, `drat`= Rear axle ratio, `wt`= Weight (in thousands of pounds). This type of plot is highly useful for identifying univariate relationships between the dependent variable (Y) and the independent variables (X's). Additionally, the matrix helps in spotting potential outliers or influential data points. While these graphs allow us to identify possible univariate relationships, they do not reveal multivariate relationships.

(i-b) **Residuals and standardized residuals**. The residuals are the deviations between the data points and the corresponding fitter values. They are defined as

$$e_i = y_i - \hat{y}_i \;\; i = 1, \ldots, n.$$

They can be viewed as estimations of the model errors, denoted as ϵ_i and represented as e_i. This implies that the residuals can serve as means to evaluate the model's assumptions concerning the errors. Standardized residuals are obtained by dividing the residuals by their estimated standard deviations $\sum_{i=1}^{n}(e_i - \bar{e})^2/(n-p)$, such that

$$\tilde{r}_i = \frac{e_i}{\sqrt{\sum_{i=1}^{n}(e_i - \bar{e})^2/(n-p-1)}}, \;\; i = 1, \ldots, n,$$

effectively normalizing them to a standard deviation of 1. This normalization assists in gauging their magnitude. If there are numerous standardized residuals with an absolute value exceeding 2, it suggests that the error distribution possesses heavier tails than a normal distribution. However, it's important to note that approximately 5% of standardized residuals are expected to exceed a magnitude of 2 even if the errors follow a normal distribution. This is due to the fact that 2 is approximately equal to 1.96, which corresponds to the 97.5-th percentile of the standard normal distribution. However, theoretically speaking, residuals do not exhibit constant variance like errors. In fact, it is trivial to show that the covariance matrix of the residuals is

$$Var(\mathbf{e}) = (\mathbf{I}_n - \mathbf{H})Var(\mathbf{y}) = \sigma^2 \left[\mathbf{I}_n - \mathbf{H}\right],$$

where

$$\mathbf{H} = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}},$$

is called the hat matrix.

In details, the variance-covariance matrix of the residual vector has two main elements of interest

$$Var\left(e_i\right) = \sigma^2\left(1 - h_{ii}\right),$$

and

$$Cov\left(e_i, e_j\right) = -\sigma^2 h_{ij} \; i \neq j,$$

where h_{ij} is the $(i,j)^{th}$ element of $\mathbf{H}$.

While the errors are assumed to be uncorrelated, the above result shows that the residuals are, in general, correlated. This also suggests a formula for the standardised residuals. For the i^{th} observation, the standardised residual is

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1 - h_{ii}}},$$

it follows that the studentized residuals have constant variance $Var(r_i) = 1$.

(i-c) Residual plots: Using standardized residuals for residual plots is logical, especially when seeking heteroscedasticity, as they are expected to display consistent variance. However, examining raw residuals can also be informative.

There are several possible residual plots. For instance in the **predictor vs. residuals plot**, the errors are assumed to be independent from the predictor variables, and therefore the residuals should look randomly scattered when plotted against every single predictor variable. Otherwise, the plot can reveal non-linearity, heteroscedasticity, autocorrelation (i.e. dependence among themselves) of residuals with neighboring values of the predictor and outliers.

(v) **Fitted values vs. residuals.** If the model is true, the correlation between the residuals and

fitted values is zero . The plot should look randomly scattered. The fitted values are, mathematically, a linear combinations of the predictors. Therefore, this plot can reveal the same kind of problems as the predictor vs. residuals plot.

(vi) **Observation order vs. residuals.** This makes sense if the observation order is informative (i.e. time). As the errors are assumed to be i.i.d., this plot should look randomly scattered as well. It often reveals autocorrelation among the residuals, but may show heteroscedasticity as well. (As long as the observation order is not a predictor in itself, there is no linearity assumption to be checked here, but strongly nonlinear patterns may still be worth investigation.)

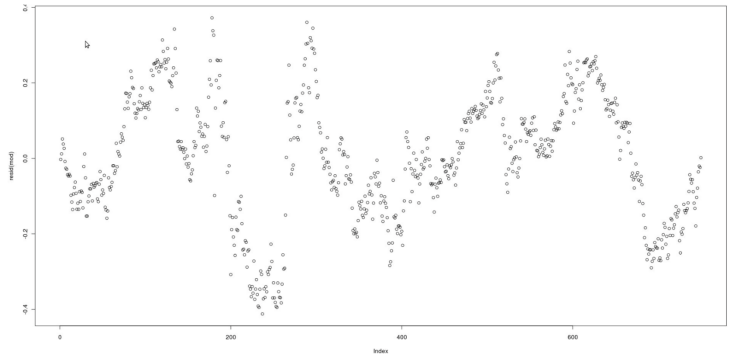


Figure 3.8: The residuals in this plot don't appear to be randomly distributed around zero; instead, they seem to follow a specific seasonal pattern. This plot could indicate the presence of potential autocorrelation in the residuals. The presence of autocorrelation violates assumption (A3) of the model.

(vii) **Normal probability plot of residuals.** The normal probability plot plots the sorted (standardized) residuals $r_{(i)}$ (denoting the i th smallest residual) against the theoretical quantiles of the normal distribution $\left(\Phi^{-1}\left(\frac{i-0.5}{n}\right)\right)$, which are the "ideal" locations of sorted realizations of a standard normal distribution. This should look roughly like a straight line. Otherwise, it indicates deviations from normality, including outliers (which can be seen at the extreme ends of the plot).

If a nonlinearity is observed in a normal probability plot, its interpretation involves examining where the plot demonstrates convex or concave characteristics. A convex curve signifies that as one progresses from left to right, the slope of the tangent line increases (as depicted in Figure **??**(a)). Conversely, if the slope decreases as one moves from left to right, the curve is concave (as illustrated in Figure **??**(b)). A curve that is convex-concave combines convexity on the left and concavity on the right, while a concave-convex curve presents concavity on the left and convexity on the right (shown in Figure **??**(c) and (d), respectively).

A normal plot demonstrating convexity, concavity, convex-concave, or concave-convex characteristics indicates left skewness, right skewness, heavier tails (in comparison to the normal distribution), or lighter tails (in comparison to the normal distribution), respectively. It's important to note that these interpretations assume the sample quantiles are on the horizontal axis, and adjustments are required if the sample quantiles are plotted on the vertical axis. The tails of a distribution refer to the regions located far from its center.

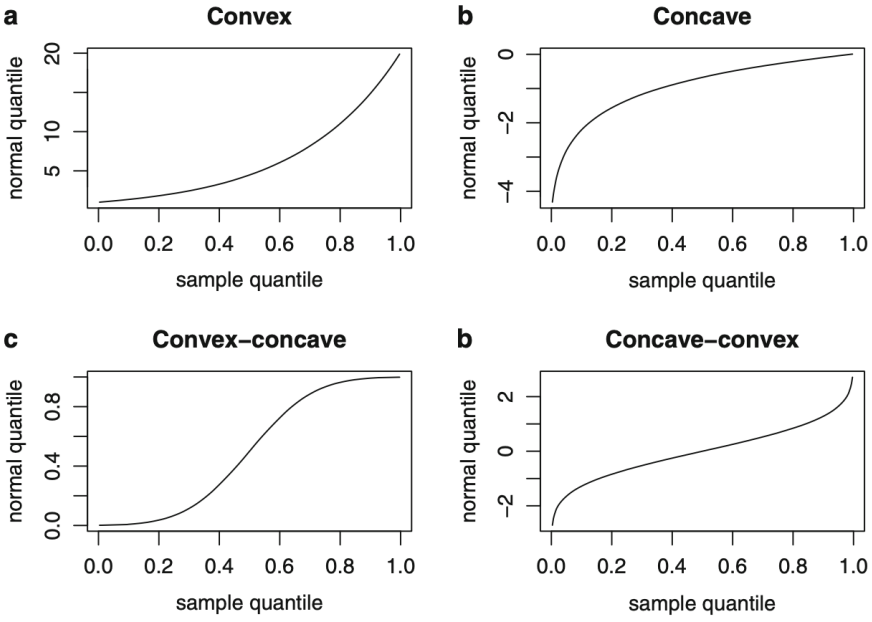


Figure 3.9: As one moves from (a) to (d), the curves are convex, concave, convex- concave, and concave-convex. Normal plots with these patterns indicate left skewness, right skewness, heavier tails than a normal distribution, and lighter tails than a normal distribution, respectively, assuming that the data are on the x-axis and the normal quantiles on the y-axis.

3.6.1 Remedies for violated model assumptions.

(i) **Non-linearity.** Occasionally, employing transformations on the predictors and/or the response variables can be beneficial. A linear model might be applicable to certain nonlinear functions derived from the observed variables. Common initial choices involve taking logarithms, square roots, squares, or exponentials. (Futher details can be found in Chapter 5: Introduction to linear regression analysis, Montgomery)

(ii) **Non-normality of the error distribution.** Robust linear regression (not treated in this course Chapter 5: Introduction to linear regression analysis, Montgomery) may help, particularly with outliers. In case of skewness, transformations could improve the situation. A possibility is to apply the same transformation to response and predictors.

(iii) **Heteroscedasticity** (as opposed to "homoscedasticity"). Weighted least squares (or iterative reweighting); transformations; sometimes (in the case that heteroscedasticity)

(iv) **Dependence of errors.** Sometimes this does not affect regression parameter estimators, but it does affect standard deviations and confidence intervals. If assumptions about the nature of dependence can be made, time series models may apply.

- Coefficient of determination, R^2. This is defined by

$$R^2 = 1 - \frac{\text{Dev}(\hat{e})}{\text{Dev}(y)} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{RSS}{TSS},$$

which is the proportion of the total variation explained by the regression model. We have seen that the R^2 is the square of the correlation between the observed and fitted values . This correlation is known as the multiple correlation coefficient.

The coefficient of determination (with a maximum value of 1) measures how well the model accounts for the data. Note, however, that it is not directly related to the model assumptions. A small value of R^2 may occur in case that assumptions are violated or that some crucial information (further predictors) is missing in the data or that the model is fine but the error variance is large, so that the predictors don't have a large explanatory or predictive strength. Violations of the model assumptions are still possible if R^2 is relatively high. If the true relationship between predictors and response is strong and monotone but slightly nonlinear, or in case of heteroscedasticity with comparably small error variances, a linear regression may still yield a relatively good fit and a high R^2.

(v) **Outliers.** Outliers are observations that, in some way, behave differently from the bulk of the data. They can for instance be extreme in the x-direction or in the y-direction. Hence we distinguish between the following:

- **Regression outliers:** these are observations that have an unusual y-value compared to other observations with similar x-values. If there are only few regression outliers these may show up in residual plots as having large residuals.

- **Leverage points:** these are observations that have unusual x-values compared to the bulk of the data. In a designed experiment they can be prevented but they are very common in observational data. Note that linear regression does not assume normality for the predictors, and therefore leverage points do not violate the model assumptions (except if they are "bad", see below). But they cause instability of the regression in the sense that small modifications of the data may lead to large changes in the least squares regression estimator.

Depending on whether the y-value is also unusual we further distinguish:

a) **Good leverage points:** if the y-value is 'in line' with other y-values. Typically, for a good leverage point the fitted value $\mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}$ is similar to the observed value y_i and omitting this observation will not change the fitted model dramatically.

b) **Bad leverage points:** if the y-value is also unusual. For a bad leverage point the fitted value $\mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}$ can be close to or very different from the observed value y_i- depending on the extend of the leverage effect. An example is presented in Figure **??**.

Hence leverage describes the potential for affecting the model fit. An intuitive way to measure the leverage of an observation $(\mathbf{x}_i, y_i)$ is to calculate how far away is $\mathbf{x}_i$ from the centre of the x-values. As this is scale dependent we should standardise this distance, which leads to the following notion of Mahalanobis Distance MD_i

$$MD_i = \sqrt{(\mathbf{x}_i - \overline{\mathbf{x}})^{\mathrm{T}}\, \hat{\Sigma}_x^{-1}\, (\mathbf{x}_i - \overline{\mathbf{x}})},$$

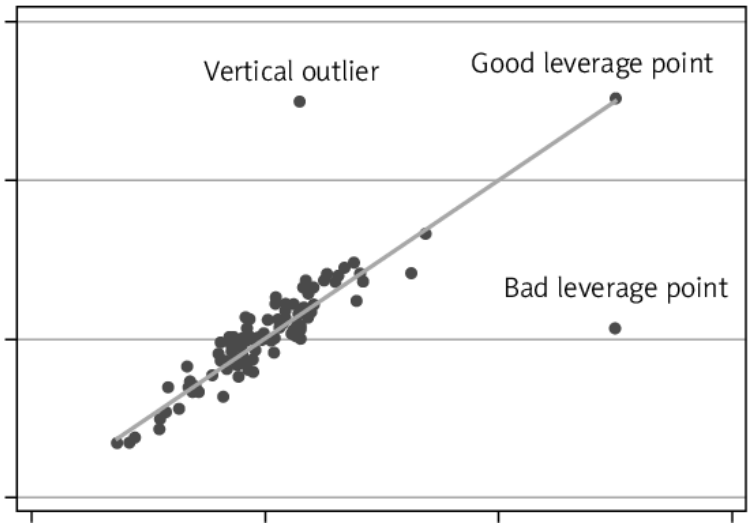


Figure 3.10: Outliers and leverage points in a simple regression analysis $Y \sim X$.

where $\hat{\Sigma}_x$ is the empirical covariance matrix of the x-observations and $\bar{\mathbf{x}}$ is the vector containing the means. It can be shown that there is a one-to-one relation with the diagonal elements of the Hat matrix:

$$MD_i^2 = (n-1)\left[h_{ii} - \frac{1}{n}\right],$$

where n is the sample size, the leverage of the i^{th} observation is therefore often just measured by h_{ii}. Under multivariate normality of the predictors (which is usually not assumed in regression), $MD_i^2 \sim \chi^2_{p-1}$, which can be used to assess whether a mahalanobis distance is unusually large. Note that leverage points can be prevented in designed experiments.

Least squares is heavily affected by outliers and it is theorefore a good idea to check, before even fitting the data, if there are any obvious unusual observations, e.g., by looking at the matrix plot (leverage points can be found in the plots of pairs of the predictor variables). However, in higher dimensions (i.e. more than 2 explanatory variables) we might not be able to spot such unusual observations by looking at 2-dimensional plots. The residual plots can help, but it has to be kept in mind that the residuals are computed from the fitted model which might itself already be affected and distorted by outliers.

The problem with influential observations such as bad leverage points is that they may have such a large influence on the least squares regression that they actually produce a smaller residual than other points and cannot be revealed by residual plots.

Another possibility of finding out about the influence of an observation is Cook's statistic. This is probably the most popular measure of influence among those proposed. Fit the model repeatedly, always omitting one observation, and see how that changes the fitted values: the change in fitted values is $\mathbf{X}\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}\right)$ where $\hat{\boldsymbol{\beta}}_{(i)}$ denotes the least squares estimator of $\boldsymbol{\beta}$ without the i^{th} observation. Cook's statistic is

$$D_i = \frac{1}{p\hat{\sigma}^2}\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}\right)^{\mathrm{T}} \mathbf{X}^{\mathrm{T}}\mathbf{X}\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}\right).$$

A large value of D_i indicates that the i^{th} observation is influential.

The numerator of D_i is just the sum of the squared differences of the fitted values with and without the i^{th} observation. It can also be shown that

$$D_i = \frac{1}{p}\left(\frac{h_{ii}}{1-h_{ii}}\right) r_i^2,$$

which means that D_i can be computed easily for all i without the need to actually fit the regression model dropping one observation at a time.

Unfortunately, neither the Mahalanobis distance nor Cook's statistic can reliably find all leverage points. A reason for this is the so-called masking effect, which means that if there is more than one leverage point (particularly at about the same location), they all together can prevent that every single one of them produces an unusual value on any of these statistics.

The only possibility to cope with this is to use robust estimators which are less sensitive even to groups of leverage points. Residual plots with residuals computed from robust estimators are more informative about outliers.

3.6.2 Multicollinearity

A serious problem that can affect a regression model is multicollinearity. Multicolinnearity implies near linear dependence among the regressors. The regressors are nothing more than the columns of the design matrix $\mathbf{X}$. The problem can we split in two scenarios:

- **Exact multicollinearity** means that there is linear dependence among the predictor variables. In that case, $\mathbf{X}^\top\mathbf{X}$ is not invertible and the least squares estimator does not exist.

- **Near multicollinearity** means that the predictors are nearly linearly dependent. This happens particularly if some of the predictors are strongly correlated (not exactly). This may be detected from the matrix plot (though sometimes the interplay of more than two variables may produce collinearity, and situations with a high number of variables and a relatively low number of data points are again, dangerous). Even though the least squares estimator can be computed in this case and it is not a violation of the model assumptions, approximate collinearity is problematic. The fact that $\mathbf{X}^\top\mathbf{X}$ is close to singularity means that some of the regression parameter estimators (particularly those belonging to strongly correlated predictors) may be very unstable and should be interpreted with care (the covariance matrix of $\hat{\beta}$, may contain some very large variance entries).

  An intuitive reason is that if two predictors measure more or less "the same thing" (i.e. are highly correlated), it cannot be clearly separated what any of them contributes to the

explanation of the response.

3.7 ANOVA- One Way- Layout

One-way analysis of variance

Suppose there are N experimental units in total and they are allocated to the treatments at random. Let y_{ij} be the j^{th} observation in the i^{th} treatment group, where $j = 1, \ldots, n_i, i = 1, \ldots, r$, where n_i is the sample size of the i^{th} group. We must have $N = \sum_i n_i$. We wish to decide whether or not the group means $\mu_1, \mu_2, \ldots, \mu_r$ are equal i.e. to test the hypothesis $H_0 : \mu_1 = \mu_2 = \cdots = \mu_r$ against H_1 : at least one pair of group means is unequal.

In this situation we might start by considering the data as observed values of random variables Y_{ij} satisfying the linear model

$$Y_{ij} = \mu_i + \epsilon_{ij} \quad (i = 1, \ldots, r; j = 1, \ldots, n_i)\,, \tag{3.13}$$

where the ϵ_{ij} are independent $\mathcal{N}(0, \sigma^2)$ r.v.'s, where we are assuming a common variance within groups. Equivalently, if we define $\alpha_i = \mu_i - \mu$, where $\mu = \sum_i n_i \mu_i / N$ is the weighted average of the μ_i, we can rewrite the model in additive form as

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}.$$

Here μ represent the 'overall mean', and α_i is an 'additional' effect of the i^{th} treatment. However, there is a problem, since the model now has $r + 1$ parameters to describe just r individual group means. We are estimating the mean of each of the r groups plus the the overall mean. Clearly, these parameters are not all uniquely defined. To overcome this problem we must impose a constraint on the parameters. The constraint that ensures the intended interpretation is the sum-to-zero parameterisation defined by $\sum_i n_i \alpha_i = 0$ (other constraints are also possible, their tratment is out of the scope of this course). In terms of the α_i, the null hypothesis can be now stated as $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_r = 0$.

Analysis: partition of the total sum of squares. A very convenient way to study this model, at least for the purpose of testing H_0, is to write each observation as a sum of deviations:

$$y_{ij} = \bar{y}_{..} + (\bar{y}_{i.} - \bar{y}_{..}) + (y_{ij} - \bar{y}_{i.})\,, \tag{3.14}$$

where $\bar{y}_{1.}, \ldots, \bar{y}_{r.}$. denote the individual group means and $\bar{y}_{..} = \sum_{i=1}^{r} n_i \bar{y}_i / \sum_{i=1}^{r} n_i$ is the overall mean. Notice the notational convention here: a dot is used to indicate a suffix over which averaging has taken place.

The three terms on the right-hand side of **??** may be thought of as the 'sample' analogues of μ, α_i and ϵ_{ij} respectively - indeed, it can be proved that the first two are the least-squares estimates

of μ and α_i. We now consider the total variation of the data about the overall mean $\bar{y}_{..}$.

$$\sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{..})^2=\sum_{i=1}^{r}\sum_{j=1}^{n_i}\{(\bar{y}_{i.}-\bar{y}_{..})+(y_{ij}-\bar{y}_{i.})\}^2.$$

Expanding out the bracket on the right-hand side here, it can be shown (see exercises) that the cross-product term vanishes to yield

$$\sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{..})^2=\sum_{i=1}^{r}n_i(\bar{y}_{i.}-\bar{y}_{..})^2+\sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i.})^2.$$

The first term on the right-hand side in some sense represents variation among the different group means: it is called the between-groups sum of squares (s.s) and denoted below by S_G :

$$S_G=\sum_{i=1}^{r}n_i(\bar{y}_{i.}-\bar{y}_{..})^2.$$

The second term is a sum of contributions representing the variation of observations within each individual group about the group means: it is called the residual or error or within-groups sum of squares (s.s.):

$$S_E=\sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2=\sum_{i=1}^{r}(n_i-1)\,s_i^2.$$

Denoting the total variation by S_T therefore, we have

$$S_T = \quad S_G \quad + S_E$$

It seems intuitively reasonable to assess the plausibility of $H_0 : \alpha_1 = \ldots = \alpha_r = 0$ by considering the value of S_G relative to that of S_E. This can be done by computing the variance ratio, or mean-square ratio

$$F=\frac{S_G/(r-1)}{S_E/(N-r)} \tag{3.15}$$

which should be approximately 1 under H_0 but larger than 1 otherwise.

The decomposition of the total variation is an algebraic result that does not depend on the assumed probability model. To formulate a test of H_0, however, we need to be able to say something about the distributions of S_E and S_G. To do this, we need to make use of the normality assumption. With a bit of albegra and using some well know theorems, it can be proved that $S_E/\sigma^2 \sim \chi^2_{(N-r)}$, $S_G/\sigma^2 \sim \chi^2_{(r-1)}$ and that S_E, S_G are independent, where χ^2 denote a chi-squared distribution. Therefore, under H_0, the mean-square ratio (**??**) has a Snedecor's F distribution with $r-1$ and

$N - r$ dregrees of freedom, we will denote this with the following symbol $F_{(r-1,N-r)}$.

The Analysis of variance (ANOVA) table includes the sum of squares, the mean square (m.s.), the mean-square ratio (m.s.r.), the degree of freedom (d.f.) and the F test:

Source of variation	Sum of squares	d.f.	m.s.	m.s.r.	Expected m.s.
Between groups	$S_G = \sum_{i=1}^{r} n_i \left(\bar{y}_i - \bar{y}_.\right)^2$	$r-1$	$\frac{S_G}{r-1}$	$F = \frac{S_G/(r-1)}{S_E/(N-r)}$	$\sigma^2 + \frac{\sum_{i=1}^{r} n_i \alpha_i^2}{r-1}$
Within groups	$S_E = \sum_{i=1}^{r} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_{i.}\right)^2$	$N-r$	$\frac{S_E}{N-r}$		σ^2
Total	$S_T = \sum_{i=1}^{r} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_{..}\right)^2$	$N-1$			

The expected m.s. column is not included in practice. To perform the test, compute the m.s.r. F and reject H_0 at level α if $F \geq F_{\alpha,(r-1,N-r)}$, the upper $100\alpha\%$ point of the F distribution.

Suppose that we want to construct a confidence interval for the difference $\mu_i - \mu_j$ or equivalently $(\alpha_i - \alpha_j)$, where μ_i is the mean of the i^{th} and μ_j is the mean of the j^{th} group. An unbiased estimator is $\bar{y}_{i.} - \bar{y}_{j.}$ with standard error given by $S\sqrt{n_i^{-1} + n_j^{-1}}$, where $S = \sqrt{\frac{S_E}{N-r}}$ is the estimate of σ. Therefore a $100(1-\alpha)\%$ confidence interval for $\mu_i - \mu_j$ is

$$\bar{y}_{i.} - \bar{y}_{j.} \pm t_{\alpha/2,(N-r)} S \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}.$$

Example

(Clarke & Kempson) In an experiment to compare four different fuel mixtures, mixtures were randomly assigned to each of 25 motor vehicles and the fuel consumption of each (in $\mathrm{km/L}$) was measured under controlled conditions. The results were as follows:

A	13.31	14.04	13.68	13.75	13.12	14.11	13.96
B	14.28	14.47	14.03	15.62	15.10		
C	15.04	14.77	15.13	15.45	14.98	15.51	
D	14.66	13.93	15.05	14.21	14.42	14.30	14.25

The calculations are most easily presented in tabular form:

Treatments	n_i	y_i	$\sum_j y_{ij}^2$	$y_{i.}^2/n_i$
A	7	95.97	1316.5907	1315.7487
B	5	73.50	1082.1346	1080.4500
C	6	90.88	1376.9344	1376.5291
D	7	100.82	1452.8760	1452.0961
	$N = 25$	$y.. = 361.17$	$S = 5228.5357$	5224.8239

Total s.s. $= S - y_{..}^2/N = 5228.5357 - (361.17)^2/25 = 5228.5357 - 5217.7508 = 10.7849.$

Between-groups s.s. $= \sum_{i=1}^{r} y_{i.}^2/n_i - y_{..}^2/N = 5224.8239 - 5217.7508 = 7.0731.$

The ANOVA table is therefore as follows:

Source of variation	s.s.	d.f.	m.s.	m.s.r.
Between groups	7.0731	3	2.3577	13.339
Within groups	3.7118	21	0.1768	
Total	10.7849	24		

The 0.1% point of F with 3 and 21 degrees of freedom is 7.94 . Since the observed value of F is greater than this, the result is significant at the 0.1% level. There is very strong evidence that mean fuel consumption differs under the four different mixtures. The four mean values are 13.71, $14.70, 15.15$ and 14.40, indicating that treatment C is the best $(\mathrm{km/L})$.

A 99% confidence interval for the difference between treatment C and A is $(15.14 - 13.71) \pm 2.831 \times 0.4204\sqrt{\frac{1}{6} + \frac{1}{7}}$, giving the interval $(0.774, 2.099)$.

Example in R

```
fuel.con <- c(13.31,14.04,13.68, 13.75, 13.12, 14.11, 13.96,
              14.28, 14.47,14.03, 15.62, 15.10,
              15.04, 14.77, 15.13, 15.45, 14.98, 15.51,
              14.66, 13.93, 15.05, 14.21, 14.42, 14.30, 14.25)

mixture <- as.factor(c(rep('A',7), rep('B', 5), rep('C', 6), rep('D', 7)))

fit <- aov(fuel.con ~ mixture)
summary(fit)
            Df Sum Sq Mean Sq F value   Pr(>F)
mixture      3  7.073  2.3577   13.34 4.29e-05 ***
Residuals   21  3.712  0.1768
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Checking model assumptions

The assumptions of the additive model (**??**) are firstly that the observations are from r independent normal populations, and secondly that the r population variances are equal. Some comments on these:

1. The F test is fairly robust to departures from normality if the sample sizes are sufficiently large (central limit effect), but not for very small samples. To check, plot the residuals

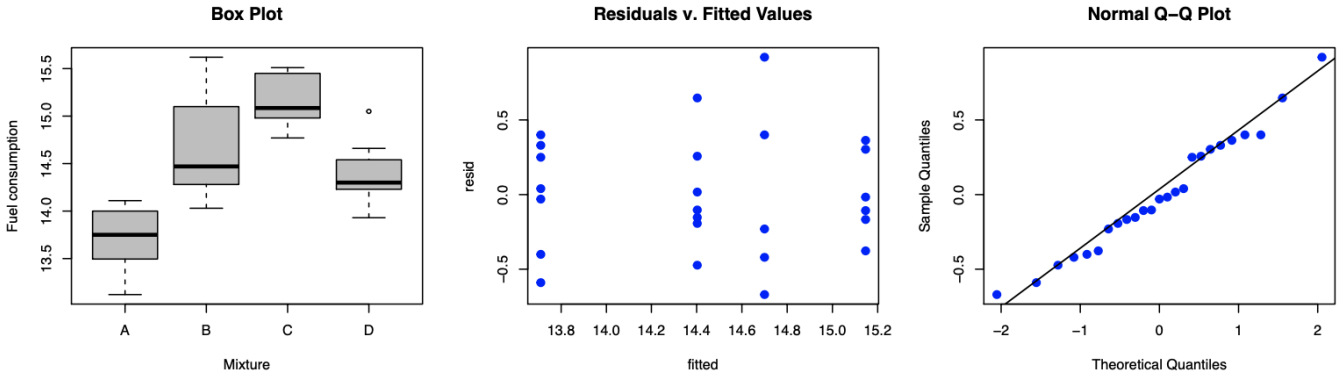


Figure 3.11: Diagnostic plots ANOVA

$r_{ij} = y_{ij} - \bar{y}_{i\cdot}$ as Q-Q plot (see earlier plots for Example). Remedies: possibly transform the data or use a different distributional model or use a distribution-free analysis (nonparametric method).

2. Independence, and homogeneity within groups, can be controlled to some extent using randomisation. However, even in this case it is useful to plot residuals against other variables in case of unexpected problems. For example, if an experiment is carried out by taking measurements over the course of a day then it may be useful to plot the residuals in time sequence - it may turn out, for instance, that observations taken in the morning are generally less than those taken in the afternoon. If such an effect is identifiable, a more sophisticated analysis may be required - for example, using a randomised blocks layout (out of the scope of this course).

3. Modest failure of the 'constant variance' assumption is not too serious if the group sizes are similar. The assumption can be checked by inspection of the within-groups sample variances $s_1^2, \ldots, s_r^2$, or using residual plots. If necessary, the hypothesis $H_0 : \sigma_1^2 = \cdots = \sigma_r^2$ can be tested formally using Bartlett's test (be careful though - this test is very sensitive to departures from normality).

If the variances appear non-constant, a possible remedy is to transform the data using a variance-stabilising transformation. Such a transformation may also result in better normality as well. For example, if the model is multiplicative: $Y_{ij} = \mu_i \epsilon_{ij}$ (so that s.d. $\propto$ group mean) then transform to an additive model by taking $\log Y_{ij}$.

3.8 Missing Data

In this section, we will discuss some fundamental concepts related to missing data. However, since the topic covered here is primarily theoretical, we will provide only essential definitions. Practical applications will be extensively addressed in the labs.

3.8.1 Missing data mechanism

Missing data is a common issue in statistical analysis, and it can cause bias or inaccurate results. The three types of missing data are missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR).

- **MCAR** happens when the probability to observe a missing observation is the same for all cases. In other words the missing data mechanism is completely unrelated to the data. Let's say you're conducting a survey about people's income and education level. You collect data from a random sample of individuals, but some respondents leave certain fields blank. In an MCAR scenario, the missing data is not related to the respondents' income or education level or any other factor. For instance, consider a situation where you have collected the following data:

Respondent	Income ($)	Education Level
1	50000	Bachelor's
2	75000	Master's
3		High School
4	60000	Doctorate
5	42000	
6		Master's

Table 3.3: Example of MCAR Data.

  In this example, the missing data points (e.g., Respondent 3's income and Respondent 5's education level) are seemingly random and don't show any particular pattern. The missing data points occur regardless of the income or education level, demonstrating that the data is missing completely at random (MCAR).

- **MAR** happens when the probability of being missing is the same only within groups. Consider a study where researchers are investigating the relationship between income and health status. They collect data from a sample of participants and ask them to provide their annual income and self-reported health status on a scale from 1 to 10. However, due to privacy concerns, some participants choose not to disclose their income.

  In this example, the missing income data is related to the health status of the participants. Those with lower health status scores (like Participant 3) might be less inclined to disclose

Participant	Income ($)	Health Status
1	45000	7
2	62000	6
3		4
4	55000	8
5	38000	5
6		9

Table 3.4: Example of MAR data.

their income. On the other hand, those with higher health status scores (like Participant 6) might be more comfortable sharing their income. The missingness of income depends on the observed variable (health status), but it's not directly related to the values of the missing data (income). This scenario demonstrates "Missing At Random" (MAR) because the probability of income being missing can be explained by other observed variables, making it possible to handle the missingness through statistical techniques that consider these relationships. MAR is more general and realistic than MCAR

- **MNAR** happens when neither MCAR nor MAR assumption holds, in this case the probability that a i-th case is missing varies for reasons that are unknown to us. Imagine a study that aims to understand the relationship between income and job satisfaction. Participants are asked to provide their annual income and rate their job satisfaction on a scale from 1 to 10. However, in this scenario, the missingness in income is not only related to participants' job satisfaction, but it's also influenced by the actual income itself. Participants with higher incomes might be less willing to disclose their income, regardless of their job satisfaction level.

Participant	Income ($)	Job Satisfaction
1	60000	8
2		6
3	48000	5
4		9
5	55000	7
6		4

Table 3.5: Example of MNAR data.

In this example, the missing income data is influenced by both the job satisfaction level and the actual income. Participants with higher incomes might choose not to disclose their income regardless of their job satisfaction, while those with lower incomes might be more inclined to share their income.

This type of missingness is referred to as "Missing Not At Random" (MNAR) because the missingness is not solely dependent on observed variables like job satisfaction; it's influ-

enced by the values of the missing data (income) itself. Handling MNAR data can be more challenging since the missingness mechanism is not fully explained by observable factors, making it necessary to carefully consider the potential biases introduced by the missing data in any analysis.

Overall, identifying the type of missingness in the data is crucial in selecting appropriate methods to handle missing data. MCAR is the easiest to handle, while MAR can be dealt with using various methods. MNAR is more challenging and requires a cautious approach to address the missingness mechanism.

3.8.2 Imputation Methods

Complete case analysis (CCA)

Complete case analysis (CCA) involves excluding all cases (rows) that have at least one missing entry. If the data is influenced by MCAR (which can be challenging to ascertain in most cases), CCA yields standard errors and confidence intervals that are accurate for the diminished subset of data, although relatively larger compared to using all available data. However, if the data is not MCAR, utilizing CCA introduces significant bias into the estimates of the regression coefficients. One straightforward method to implement CCA is by utilizing the `na.omit` function in R.

Mean Imputation

Mean imputation (MI) involves replacing the missing entries of a random variable X_i (the i-th column in our dataset) with its sample mean computed from the observed values. It is important to recognize that through MI, we alter the probabilistic characteristics of the underlying distribution. Mean imputation offers a quick and straightforward approach to handle missing data; however, it tends to underestimate the variance of the imputed random variable. Additionally, it disrupts the interrelations between random variables and introduces bias into all estimates except for the sample mean.

Regression Imputation

Regression imputation (RI) integrates information from other variables to create more informed imputed values. The initial stage entails constructing a model based on the available data. Subsequently, predictions for the incomplete cases are computed using the established model, and these predictions are used to substitute the missing data.

3.9 Exercises

1. Suppose the model for simple linear regression is written as

$$Y_i = \alpha_0 + \alpha_1 (x_i - \bar{x}) + e_i \quad (i = 1, \ldots, N)$$

where $\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$.

(a) Obtain the least squares estimators of α_0 and α_1. (You may obtain these from the matrix form of the normal equations.)

(b) Obtain the covariance matrix of these least squares estimators under the usual assumptions about the errors.

(c) Show that

$$\mathrm{RSS} = C_{YY} - \frac{C_{xY}^2}{C_{xx}}$$

where $C_{xx} = \sum_{i=1}^{N} (x_i - \bar{x})^2$, $C_{xY} = \sum_{i=1}^{N} (x_i - \bar{x}) \left(Y_i - \bar{Y}\right)$ and $C_{YY} = \sum_{i=1}^{N} \left(Y_i - \bar{Y}\right)^2$.

(d) Refer to R output below (Figure **??**). If x denotes log depth and y denotes log flow, you are given that $\bar{x} = -0.88686, \bar{y} = 0.21475, C_{xx} = 1.1482, C_{xy} = 3.1738, C_{yy} = 9.3516$ using the notation above.

State how the model defined above is related to the model fitted in the R output. Use the algebraic results obtained above and a hand calculator to verify the numerical results for the following in the R output:

- the least squares estimates of β_0 and β_1,
- the residual standard error,
- estimated covariance matrix of $\hat{\beta}_0$ and $\hat{\beta}_1$.

2. The data can be obtained as a text file treevol.dat from the course webpage.

Since the volume of a cylinder is a product of the height, the diameter to the square and a constant, it could actually make sense to assume a multiplicative model, and fitting the log VOL from log HT and log D16 could make sense. Use R or any other statistics software to fit such a model. Discuss the results and compare them to the models in handout 3 .

Compute a predicted value and 95% prediction and confidence intervals for VOL for a tree with D16=10 and HT=100 from the log-transformed model fitted here (you need to take into account the transformation for this). Compare the prediction interval to the one that you get from a model that fits VOL as a linear function of D16 and HT. Here are some useful R-commands:

```
treevol <- read.table("treevol.dat",header=TRUE)
logHT <- log(treevol$HT)
logD16 <- log(treevol$D16)
logVOL <- log(treevol$VOL)
pairs(cbind(logVOL,logHT,logD16))
lmtreelog <- lm(logVOL ~ logHT + logD16)

newtree <- data.frame(HT=100,D16=10)
lognewtree <- data.frame(logHT=log(newtree$HT),logD16=log(newtree$D16))
predict(lmtreelog,lognewtree,interval=c("confidence"),level=0.95)
predict(lmtreelog,lognewtree,interval=c("prediction"),level=0.95)
```

Figure 3.12: R output Ex 4

3. Explain in your own words and understandable to a non-statistician (with some solid school background in maths), what the model assumption of i.i.d. error terms in the general linear model means. (Don't assume that the reader knows what an "error term" is.)

4. The matrix formulation of the multiple linear model is given

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

You may assume that $\mathbf{y}$ and $\boldsymbol{\epsilon}$ have multivariate normal distributions and that the components of the vector $\boldsymbol{\epsilon}$ are independently and identically distributed (i.i.d.).

   (a) Carefully define each of the terms $\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \boldsymbol{\epsilon}$ appearing in the model.

   (b) Given that $\mathbf{X}$ is a $n \times (p+1)$ matrix, state the dimensions of $\mathbf{y}$, $\boldsymbol{\beta}$ and ϵ. Explain what the values n and p refer to.

   (c) Define the hat matrix $\mathbf{H}$ for an associated general linear model and show that it has the property $\mathbf{H}^\top = \mathbf{H}$. Clearly state any assumptions about the matrix $\mathbf{X}$ and state the matrix algebra results which are being used.

   (d) Define the terms outlier and influential observation. Explain the term leverage with respect to a multiple linear regression. How this can be used to identify influential observations.

5. Suppose we are estimating the regression coefficients in a linear regression model by minimizing

$$\min_{\boldsymbol{\beta}\in\Theta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq s, \tag{3.16}$$

where y_i is the response variable for the i-th observation, x_{ij} is the value of the j-th predictor for the i-th observation, β_0 is the intercept of the model, β_j are the coefficients of the

predictors, $\boldsymbol{\beta}^\top = (\beta_0, \beta_1, \ldots, \beta_p) \in \mathbb{R}^{p+1}$ is the coefficient vector including the intercept, Θ is its parametric space, $\lambda \geq 0$ is the regularization parameter, n denotes the number of observations, and p the number of predictors (excluding the intercept). Finally, the Residual Sum of Squares (RSS) is given by $\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}\right)^2$.

(a) Discuss the role of the constraint $\sum_{j=1}^{p} |\beta_j| \leq s$.

(b) Discuss the advantages and disadvantages of optimization in Equation **??**.

(c) Provide a geometric interpretation and draw the figure associated with the optimization in Equation **??**.

(d) Discuss what happens to the parameters and RSS when s increases.

(e) Discuss what happens to the parameters and RSS when s decreases.

6. Suppose we are estimating the regression coefficients in a linear regression model by minimizing the Ridge Regression cost function. The cost function is defined as:

$$\min_{\boldsymbol{\beta} \in \Theta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}, \tag{3.17}$$

where y_i is the response variable for the i-th observation, x_{ij} is the value of the j-th predictor for the i-th observation, β_0 is the intercept of the model, β_j are the coefficients of the predictors, $\boldsymbol{\beta}^\top = (\beta_0, \beta_1, \ldots, \beta_p) \in \mathbb{R}^{p+1}$ is the coefficient vector including the intercept, Θ is its parametric space, $\lambda \geq 0$ is the regularization parameter, n denotes the number of observations, and p the number of predictors (excluding the intercept). Finally, the Residual Sum of Squares (RSS) is given by $\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}\right)^2$.

(a) Discuss the role of the constraint $\sum_{j=1}^{p} \beta_j^2$.

(b) Discuss the advantages and disadvantages of optimization in Equation (**??**).

(c) Provide a geometric interpretation and draw the figure associated with the optimization in Equation (**??**).

(d) Discuss what happens to the parameters and RSS when λ increases.

(e) Discuss what happens to the parameters and RSS when λ decreases.

7. This exercise considers hospital expenditures data provided by the U.S. Agency for Healthcare Research and Quality (AHRQ) Filename is "HospitalCosts"

(a) Produce a scatterplot, correlation, and linear regression of LNTOTCHG on AGE. Is AGE a significant predictor of LNTOTCHG?

(b) You are concerned that newborns follow a different pattern than other ages do. Create a binary variable that indicates whether AGE equals zero. Run a regression using this binary variable and AGE as explanatory variables. Is the binary variable statistically significant?

(c) Now examine the sex effect, using the binary variable FEMALE, which is one if the patient is female and zero otherwise. Run a regression using AGE and FEMALE as explanatory variables. Run a second regression running the two variables with an interaction term. Comment on whether the gender effect is important in either model.

(d) Now consider the type of admission, APRDRG, an acronym for " all patient refined diagnostic related group." This is a categorical explanatory variable that provides information on the type of hospital admission. There are several hundred levels of this category. For example, level 640 represents admission for a normal newborn, with neonatal weight greater than or equal to 2.5 kilograms. As another example, level 225 represents admission resulting in an appendectomy.

8. This exercise considers state of Wisconsin lottery sales data, Filename is "WiscLottery". You decide to examine the relationship between SALES (y) and all eight explanatory variables (PERPERHH, MEDSCHYR, MEDHVL, PRCRENT, PRC55P, HHMEDAGE, MEDINC, and POP).

(a) Fit a regression model of SALES on all eight explanatory variables.

(b) Find R^2. Use it to calculate the correlation coefficient between the observed and fitted values.

(c) Test whether POP, MEDSCHYR, and MEDHVL are jointly important explanatory variables for understanding SALES.

After the preliminary analysis , you decide to examine the relationship between SALES(y) and POP, MEDSCHYR, and MEDHVL.

- Fit a regression model of SALES on these three explanatory variables.
- Has the coefficient of determination decreased from the eight-variable regression model to the three-variable model? Does this mean that the model is not improved or does it provide little information? Explain your response.
- To state formally whether one should use the three- or eight-variable model, use a partial F -test. State your null and alternative hypotheses, decision-making criterion, and decision-making rules.

```
> # Fitting regression model, logflow on logdepth
> flow.lm<-lm(logflow~logdepth, flow)
> # Results
> summary(flow.lm)
Call:
lm(formula = logflow ~ logdepth, data = flow)
Residuals:
       Min         1Q     Median         3Q        Max
-0.3870586 -0.1082902 -0.0009169  0.0743591  0.5432562

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.6661     0.2383   11.19 3.65e-06 ***
logdepth      2.7641     0.2510   11.01 4.12e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.269 on 8 degrees of freedom
Multiple R-Squared: 0.9381,     Adjusted R-squared: 0.9304
F-statistic: 121.2 on 1 and 8 DF,  p-value: 4.118e-06

> # Estimate the covariance matrix of parameters estimates
> matrix.flow<-summary(flow.lm)
> matrix.flow$sigma^2*matrix.flow$cov.unscaled
            (Intercept)   logdepth
(Intercept)  0.05680066 0.05588818
logdepth     0.05588818 0.06301811
> # Analysis of variance
> anova(flow.lm)
Analysis of Variance Table

Response: logflow
          Df Sum Sq Mean Sq F value    Pr(>F)
logdepth   1 8.7727  8.7727  121.24 4.118e-06 ***
Residuals  8 0.5789  0.0724
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> # Fitted values and residuals
> cbind(fitted(flow.lm),resid(flow.lm))
         [,1]        [,2]
1  -0.3158303 -0.13672644
2  -0.7555056 -0.38705862
3  -0.8525025  0.54325623
4   0.2682553  0.01466550
5  -0.7555056  0.03601440
6   0.2016465 -0.28068973
7   1.9075597  0.08714062
8   1.7962375 -0.02298146
9   0.5197130  0.16287871
10  0.1333930 -0.01649921
```

Figure 3.13: R output Ex 1

Chapter 4

Variable Selection and Regularization

4.1 Principal Components Regression

There is another alternative strategy to standard least squares (LS) regression that addresses scenarios where multicollinearity among predictors affects the stability and interpretability of a standard linear regression model. The idea is to reduce dimensionality before applying LS, which involves transforming the original predictors into a new set of orthogonal components through Principal Components Regression (PCR).

Principal Components Regression: PCR simplifies the predictor space by utilizing q transformed variables which are linear combinations of the original p predictors, where $q < p$. These transformations are defined as

$$Z_k = \sum_{j=1}^{p} \phi_{jk} X_j,$$

for some constants $\phi_{1k}, \phi_{2k}, \ldots, \phi_{pk}$, and $k = 1, \ldots, q$. The transformed variables are orthogonal components that focus on retaining the components that explain the most variance, which often leads to more reliable predictions in the presence of high-dimensional data.

Model Formulation: A LS regression model is then fitted using these transformed variables:

$$y_i = \theta_0 + \sum_{k=1}^{q} \theta_k z_{ik}, \quad i = 1, \ldots, n.$$

Instead of estimating $p + 1$ coefficients, we estimate only $q + 1$. This reduces the dimensionality and simplifies the model by focusing on the most significant aspects of the data.

Trade-offs: However, the convenience of PCR comes with a trade-off. By prioritizing variance over the direct relationship with the response variable, PCR might exclude components that have less variance but more predictive power regarding the outcome. This results in a loss of some explanatory power, as the selected principal components may not fully capture the nuances of how specific predictors influence the response. Consequently, PCR can obscure the interpretative link

between original predictors and the response, making it less useful for understanding the exact nature of these relationships.

Interestingly, with some re-ordering, the relationship between the transformed predictors and the response can be expressed as:

$$\sum_{k=1}^{q} \theta_k z_{ik} = \sum_{j=1}^{p} \left(\sum_{k=1}^{q} \theta_k \phi_{jk} \right) x_{ij} = \sum_{j=1}^{p} \beta_j x_{ij}$$

where $\beta_j = (\sum_{k=1}^{q} \theta_k \phi_{jk})$, and this is essentially a constraint. Like penalised regression, the coefficients are constrained, but now the constraint is quite different—increasing bias but decreasing variance, yielding benefits for settings that are not "least-squares friendly".

Desired Properties of Transformations: In determining the ϕ constants, it is crucial to consider what properties the transformed variables Z should have. Ideally, if the new variables Z are uncorrelated, this would provide a solution to the problem of multicollinearity. Furthermore, if the new variables Z capture most of the variability of the X's, and assuming that this variability is predictive of the response, the Z variables will be reliable predictors of the response. This is precisely what PCR aims to achieve.

4.1.1 Principal Component Analysis (PCA)

PCR utilises Principal Component Analysis (PCA) of the data matrix. Now, we wish to change the perspective: while maintaining the same information of the original p covariates, our aim is to create new *features* that encapsulate all the information of the original variables, but in a reduced number $M < p$.

Let $Z_1, \ldots, Z_M$ be linear combinations of the p predictors, where $M < p$. If this were not the case, considering a transformation of the covariates would make little sense; in such a scenario, it would be more logical to directly use the original covariates. Then

$$Z_m = \phi_{1m} X_1 + \phi_{2m} X_2 + \cdots + \phi_{pm} X_p$$

for $m = 1, \ldots, M$ and considering that $\phi_{1m}, \ldots, \phi_{pm}$ are constants (are constant in our regression problem, i.e. they must not be estimated via OLS).

Assuming we have observed the values of $\phi_{1m}, \ldots, \phi_{pm}$, we can estimate the model by now considering $z_1, \ldots, z_M$. Therefore, we can write

$$y_i = \theta_0 + \theta_1 z_{i1} + \cdots + \theta_M z_{iM} + \epsilon_i \quad i = 1, \ldots, n, \tag{4.1}$$

observing the Equation **??** we can notice that now we are no longer using the X's and that in

order to emphasize the change of variable we have used another notation for the coefficients which are now $\theta_0, \theta_1, \ldots, \theta_M$.

It is well-known in the literature that if the constants $\phi_{1m}, \ldots, \phi_{pm}$ are chosen appropriately, then the formula **??** performs better than the OLS method applied to $X_1, \ldots, X_p$. However, this is particularly true if there is a strong linear relationship among the p covariates. Indeed, if the p covariates are independent, it can be shown that estimating formula **??** or applying the OLS method to $X_1, \ldots, X_p$ is completely equivalent.

But why are we discussing dimensionality reduction? The reason is that while previously our problem was to estimate $p+1$ coefficients, now we need to estimate 'only' $M+1$. The advantage of this approach is particularly evident in high-dimensional contexts, namely when $p \gg n$ or p is very close to n. In these situations, there is a risk of overfitting the model, leading to imprecise and highly variable estimates. By shifting from $p+1$ to $M+1$, we are reducing the dimensionality while preserving the information of the original covariates.

As previously discussed, PCA (Principal Component Analysis) is a data reduction technique. When analyzing large datasets in which one or more covariates are strongly correlated, instead of discarding these correlated covariates due to potential multicollinearity issues, PCA allows for the condensation of the information contained in p covariates into $M < p$ new features.

Now, we are introducing PCA (Principal Component Analysis) in a regression context, but it can also be useful in visualization scenarios. Indeed, in data visualization, representing more than 3 dimensions becomes challenging. Techniques to enhance visual dimensionality are well-known (such as using colors, different shapes, various sizes, etc.), but sometimes they are not sufficient. In these cases, PCA can serve as a valuable tool to compress dimensions and facilitate easier visualization. If we think about, this is the same mechanism by which, when we have to handle a large file, we prefer to compress the file before sending it via email.

For example, take the first principal component based on the set $X_1, \ldots, X_p$; this can be written as

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \cdots + \phi_{p1}X_p,$$

where $(\phi_{11}, \ldots, \phi_{1p})$ are the *loadings* of the first principal component. In order to uniquely identify Z_1, we must impose the following constraint: $\sum_{j=1}^{p} \phi_{j1}^2 = 1$. This constraint limits the variance of Z_1, preventing it from growing indefinitely.

Given that $\mathbf{X}$ is an $n \times p$ matrix, the first step before calculating the principal components is to center the data matrix $\mathbf{X}$. This step is reasonable since we are constructing the principal components to capture as much information as possible. Since variability is synonymous with information in data, we are interested in variance. Let

$$z_1 = \phi_{11}x_1 + \phi_{21}x_2 + \cdots + \phi_{p1}x_p,$$

the first principal component based on the sample observations. Our goal is to identify the coefficients $\phi_{11}, \ldots, \phi_{p1}$ (the loadings) such that they result in the maximum possible variance for Z_1, given that $\sum_{j=1}^{p} \phi_{j1}^2 = 1$. Therefore

$$\text{maximize}_{\phi_{11},\ldots,\phi_{p1}} \left[n^{-1} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{j1} x_i j \right)^2 \right] \quad \text{s.t.} \quad \sum_{j=1}^{p} \phi_{j1}^2 = 1 \tag{4.2}$$

in the **??** we can write the objective function as $n^{-1} \sum_{i=1}^{n} z_{i1}^2$ because the X's have been centered and therefore have a mean equal to zero. We refer to $z_{11}, \ldots, z_{n1}$ as the *scores* of the first principal component.

Upon determining Z_1, our next step is to compute Z_2, which represents the second principal component. This component is derived as a linear combination of $X_1, \ldots, X_p$ that maximizes variance under specific conditions: the loadings must adhere to the constraint $\sum_{j=1}^{p} \phi_{j2}^2 = 1$. Additionally, it is imperative that Z_2 is orthogonal to Z_1, ensuring that the two components are uncorrelated.

The requirement for Z_2 to be uncorrelated with Z_1 makes perfect sense, as we do not want to risk incorporating redundant information into Z_2 that has already been captured in Z_1. We can express the second principal component as follows:

$$z_{i2} = \phi_{12} x_{i1} + \phi_{22} x_{i2} + \cdots + \phi_{p2} x_{ip}$$

where $\phi_{12}, \ldots, \phi_{p2}$ are the *loadings* of the second principal component. This same reasoning is then extended to Z_3, which, in addition to the usual constraint on the *loadings*, should also be uncorrelated with both Z_1 and Z_2.

4.2 Ridge regression

In multiple regression it is shown that parameter estimates based on minimum residual sum of squares have a high probability of being unsatisfactory, if not incorrect, if the prediction vectors are not orthogonal. A way to reduce the variance of $\hat{\boldsymbol{\beta}}^{\text{OLS}}$ (i.e., the parameters obtained via the ordinary least squares method) is to shrink some of the estimated coefficients towards zero. Ridge Regression (Hoerl and Kennard, 1970) solves the following optimization problem:

$$\left(\hat{\mu}_\lambda^{\text{R}}, \hat{\boldsymbol{\beta}}_\lambda^{\text{R}} \right) = \underset{(\mu, \boldsymbol{\beta}) \in \mathbb{R} \times \mathbb{R}^p}{\arg\min} \left\{ \|Y - \mu \mathbf{1} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2 \right\}$$

Where $\mathbf{1} = (1, \ldots, 1)^T \in \mathbb{R}^n$ and $\boldsymbol{\beta} \in \mathbb{R}^p$ is a vector of dimension p, $\mathbf{X} \in \mathbb{R}^{n \times p}$ is our data matrix and λ is a non-negative scalar. The first thing we notice is that the optimization problem in Ridge is very similar to that of linear regression. The difference is the addition of an extra

penalty term $\|\boldsymbol{\beta}\|_2^2$ which geometrically can be interpreted as a p-sphere. The parameter $\lambda \geq 0$, controls the degree of penalty towards zero of the parameters. With $\lambda \equiv 0$ we return to a classic linear regression problem (unpenalized estimation problem) while a value of λ tending to infinity would force all coefficients to assume very small values but never exactly zero. λ is often called the *tuning parameter* or *regularization parameter*. In the optimization problem, we explicitly included a non-penalized intercept term.

Consider the case where we are working with a variable **temperature** which may be in Kelvin or degrees Celsius, we know that the values of the parameters would not change. However, $\mathbf{X}\hat{\beta}$ is not invariant to scale transformations of the variables, hence it is *good and correct practice* to center each column of $\mathbf{X}$ (making them orthogonal to the intercept) and then scale them to have an ℓ_2 norm equal to $\sqrt{n}$.

It is trivial to show that after having standardized the data in the design matrix $\mathbf{X}$, $\hat{\mu}_\lambda^{\mathrm{R}} = \bar{Y} := \sum_{i=1}^n Y_i/n$, it is straightforward to remember that $\sum_{i=1}^n Y_i = 0$ by replacing Y_i with $Y_i - \bar{Y}$ thus removing μ from the objective function. Ridge estimates have the following form:

$$\hat{\boldsymbol{\beta}}_\lambda^{\mathrm{R}} = \left(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}\right)^{-1} \mathbf{X}^\top \mathbf{Y}$$

The addition of the term λ stabilizes the inversion of $\mathbf{X}^T\mathbf{X}$. Remember indeed that when $\mathbf{X}$ does not have full rank, the estimates obtained with the least squares method are unstable.

An important theorem now comes into play.

Theorem. Assume that $\mathbf{X}$ is a full rank matrix. Let $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ be the estimates obtained with the ordinary least squares method (OLS), $\hat{\boldsymbol{\beta}}^R$ those obtained with the Ridge Regression method and $\boldsymbol{\beta}^0$ the true parameter vector. For sufficiently small λ,

$$\mathbb{E}\left(\hat{\boldsymbol{\beta}}^{\mathrm{OLS}} - \boldsymbol{\beta}^0\right)\left(\hat{\boldsymbol{\beta}}^{\mathrm{OLS}} - \boldsymbol{\beta}^0\right)^\top - \mathbb{E}\left(\hat{\boldsymbol{\beta}}_\lambda^{\mathrm{R}} - \boldsymbol{\beta}^0\right)\left(\hat{\boldsymbol{\beta}}_\lambda^{\mathrm{R}} - \boldsymbol{\beta}^0\right)^\top$$

is positive definite.

Proof. First, let us calculate the bias of $\hat{\boldsymbol{\beta}}_\lambda^{\mathrm{R}}$. Ignoring for the moment the subscript λ and the subscript R for convenience.

$$\begin{aligned}
\mathbb{E}(\hat{\boldsymbol{\beta}}) - \boldsymbol{\beta}^0 &= \left(\mathbf{X}^\top X + \lambda \mathbf{I}\right)^{-1} \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta}^0 - \boldsymbol{\beta}^0 \\
&= \left(\mathbf{X}^\top X + \lambda \mathbf{I}\right)^{-1} \left(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I} - \lambda \mathbf{I}\right) \boldsymbol{\beta}^0 - \boldsymbol{\beta}^0 \\
&= -\lambda \left(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}\right)^{-1} \boldsymbol{\beta}^0.
\end{aligned}$$

Now let's think about the variance of $\hat{\boldsymbol{\beta}}$.

$$\begin{aligned}
\mathrm{Var}(\hat{\boldsymbol{\beta}}) &= \mathbb{E}\left\{\left(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}\right)^{-1} \mathbf{X}^\top \boldsymbol{\varepsilon}\right\}\left\{\left(\mathbf{X}^\top X + \lambda \mathbf{I}\right)^{-1} \mathbf{X}^\top \boldsymbol{\varepsilon}\right\}^\top \\
&= \sigma^2 \left(\mathbf{X}^\top X + \lambda \mathbf{I}\right)^{-1} \mathbf{X}^\top \mathbf{X} \left(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}\right)^{-1}.
\end{aligned}$$

So,

$$\mathbb{E}\left(\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}-\boldsymbol{\beta}^{0}\right)\left(\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}-\boldsymbol{\beta}^{0}\right)^{\top}-\mathbb{E}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^{0}\right)\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^{0}\right)^{\top}$$

is equal to

$$\sigma^{2}\left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}-\sigma^{2}\left(\mathbf{X}^{\top}\mathbf{X}+\lambda\mathbf{I}\right)^{-1}\mathbf{X}^{\top}\mathbf{X}\left(\mathbf{X}^{\top}\mathbf{X}+\lambda\mathbf{I}\right)^{-1}-\lambda^{2}\left(\mathbf{X}^{\top}\mathbf{X}+\lambda I\right)^{-1}\boldsymbol{\beta}^{0}\boldsymbol{\beta}^{0^{\top}}\left(\mathbf{X}^{\top}\mathbf{X}+\lambda\mathbf{I}\right)^{-1}$$

After some simplifications, we have that

$$\lambda\left(\mathbf{X}^{\top}\mathbf{X}+\lambda\mathbf{I}\right)^{-1}\left[\sigma^{2}\left\{2\mathbf{I}+\lambda\left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}\right\}-\lambda\boldsymbol{\beta}^{0}\boldsymbol{\beta}^{0^{\top}}\right]\left(\mathbf{X}^{\top}\mathbf{X}+\lambda\mathbf{I}\right)^{-1}.$$

Therefore, we have that

$$\mathbb{E}\left(\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}-\boldsymbol{\beta}^{0}\right)\left(\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}-\boldsymbol{\beta}^{0}\right)^{\top}-\mathbb{E}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^{0}\right)\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^{0}\right)$$

is positive definite for $\lambda>0$ if and only if

$$\sigma^{2}\left\{2\mathbf{I}+\lambda\left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}\right\}-\lambda\boldsymbol{\beta}^{0}\boldsymbol{\beta}^{0^{\top}}$$

is positive definite, which turns out to be true for sufficiently small values of $\lambda>0$ (we can take $0<\lambda<2\sigma^{2}/\left\|\boldsymbol{\beta}^{0}\right\|_{2}^{2}\Big)$. The theorem therefore assures us that $\hat{\boldsymbol{\beta}}_{\lambda}^{\mathrm{R}}$ performs better than $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ provided that λ is chosen appropriately. In order to be able to use ridge regression efficiently, we must define a way to select a reasonable value for λ (this will be the subject of further studies). What this theorem does not tell us is when we expect Ridge to perform well. To discuss this point, we need to explore the relationship between ridge regression and SVD.

4.2.1 Ridge and Singular Value Decomposition (optional reading)

The Singular Value Decomposition (SVD) is a generalization of the eigenvalue-based decomposition of a square matrix (square meaning that the number of rows is equal to the number of columns). The SVD allows this idea to be generalized and to factorize any $\mathbf{X}\in\mathbb{R}^{n\times p}$, even in contexts where one does not work with square matrices. Thus, we have

$$\mathbf{X}=\mathbf{U}\mathbf{D}\mathbf{V}^{\top}$$

where $\mathbf{U}\in\mathbb{R}^{n\times n}$ and $\mathbf{V}\in\mathbb{R}^{p\times p}$ are orthogonal matrices and $\mathbf{D}\in\mathbb{R}^{n\times p}$ has diagonal elements such that $\mathbf{D}_{11}\geq\mathbf{D}_{22}\geq\cdots\geq\mathbf{D}_{mm}\geq 0$, where $m:=\min(n,p)$, and all other elements of $\mathbf{D}$ are zero. The r-th columns of $\mathbf{U}$ and $\mathbf{V}$ are known as the r-th left and right singular vectors of $\mathbf{X}$, and $\mathbf{D}_{rr}$ is the r-th singular value.

When $n>p$, we can replace $\mathbf{U}$ with its first p columns and $\mathbf{D}$ with its first p rows to obtain another version of the SVD (sometimes known as thin SVD). Then $\mathbf{X}=\mathbf{U}\mathbf{D}\mathbf{V}^{\top}$ where $\mathbf{U}\in\mathbb{R}^{n\times p}$

has orthonormal columns (but is no longer square) and $\mathbf{D}$ is a square diagonal matrix. There is an equivalent version when $p > n$.

Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ be our matrix of predictors and suppose $n \geq p$. Using the (thin) SVD, we can write the fitted values of ridge regression in the following way:

$$\begin{aligned}
\mathbf{X}\hat{\beta}_\lambda^{\mathrm{R}} &= \mathbf{X}\left(\mathbf{X}^\top\mathbf{X} + \lambda I\right)^{-1}\mathbf{X}^\top Y \\
&= \mathbf{U}\mathbf{D}\mathbf{V}^\top\left(\mathbf{V}\mathbf{D}^2\mathbf{V}^\top + \lambda I\right)^{-1}\mathbf{V}\mathbf{D}\mathbf{U}^\top Y \\
&= \mathbf{U}\mathbf{D}\left(\mathbf{D}^2 + \lambda I\right)^{-1}\mathbf{D}\mathbf{U}^\top Y \\
&= \sum_{j=1}^{p} \mathbf{U}_j \frac{\mathbf{D}_{jj}^2}{\mathbf{D}_{jj}^2 + \lambda}\mathbf{U}_j^\top Y.
\end{aligned}$$

where $\mathbf{U}_j$ is the j-th column of U. For comparison, the fitted values from ordinary least squares regression (OLS) (when $\mathbf{X}$ has full rank) are

$$\mathbf{X}\hat{\beta}^{\mathrm{OLS}} = \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top Y = \mathbf{U}\mathbf{U}^\top Y$$

Both OLS regression and ridge regression compute the coordinates of Y with respect to the columns of $\mathbf{U}$. Ridge regression then shrinks these coordinates by the factors $\mathbf{D}_{jj}^2/\left(\mathbf{D}_{jj}^2 + \lambda\right)$; if $\mathbf{D}_{jj}$ is small, the amount of shrinkage will be greater. Adding another layer, observe that SVD is intimately connected to Principal Component Analysis (PCA). Consider $v \in \mathbb{R}^p$ with $|v|_2 = 1$. Since the columns of $\mathbf{X}$ have had their means subtracted, the sample variance of $\mathbf{X}v \in \mathbb{R}^n$ is

$$\frac{1}{n}v^\top\mathbf{X}^\top\mathbf{X}v = \frac{1}{n}v^\top V D^2 V^\top v$$

Writing $a = \mathbf{V}^\top v$, thus $|a|_2 = 1$, we have

$$\frac{1}{n}v^\top\mathbf{V}\mathbf{D}^2V^\top v = \frac{1}{n}a^\top\mathbf{D}^2 a = \frac{1}{n}\sum_j a_j^2\mathbf{D}_{jj}^2 \leq \frac{1}{n}\mathbf{D}_{11}\sum_j a_j^2 = \frac{1}{n}\mathbf{D}_{11}^2.$$

Since $|\mathbf{X}\mathbf{V}_1|\,2^2/n = \mathbf{D}_{11}^2/n$, $\mathbf{V}_1$ determines the linear combination of columns of $\mathbf{X}$ that has the greatest sample variance, when the coefficients of the linear combination are constrained to have an ℓ_2 norm of 1. $\mathbf{X}\mathbf{V}_1 = \mathbf{D}_{11}U_1$ is known as the first principal component of $\mathbf{X}$. The subsequent principal components $\mathbf{D}_{22}\mathbf{U}_2, \ldots, \mathbf{D}_{pp}\mathbf{U}_p$ have a maximum variance of $\mathbf{D}_{jj}^2/n$, provided they are orthogonal to all those preceding - see the list of examples 1 for details.

Returning to ridge regression, we see that it further reduces Y in the smaller principal components of $\mathbf{X}$. Therefore, it will perform well when most of the signal is found in the larger principal components of $\mathbf{X}$.

4.3 Variable Selection

Considering again the multiple regression model in the form

$$Y_i = \beta_0, \beta_1 x_{i1}, \ldots, \beta_p x_{im} + \epsilon_i, \text{when} i = 1, \ldots, n$$

In many applied fields such as Economics, Biology, Psychology, and Social Sciences, there is often an accumulation of large datasets, which may not all be informative. Specifically, it is common to handle datasets with many columns (where the number of variables p is very large) and there arises a need to select only a subset of relevant variables. The reasons for this might include:

1. Theoretical reasons might suggest that only a few covariates explain the dependent variable Y. For example, an advertising company, having collected a myriad of consumer data, may want to focus on a few important characteristics to optimize and channel advertising expenditures more efficiently.

2. Often, situations arise where $p > n$. In these cases, the problem of variable selection takes on a completely different aspect. It is not feasible to estimate a regression model (or any model in general) when the number of parameters to be estimated exceeds the number of observations. For instance, to estimate a mean, at least two observations are required. Estimating a mean with just one observation does not make sense and is not even possible. This reasoning applies to all statistics. In general, to estimate p parameters, at least the same number of observations is needed. In other words, if the number of regression parameters p is large compared to the number of observations n, $\mathbf{X}^\top\mathbf{X}$ is often close to collinearity and the estimation of the parameters can be very unstable. Getting rid of some variables (and therefore some parameters) can improve the situation. (A rule of thumb is that 10-20 observations are needed per estimated regression parameter in order to estimate them accurately enough.)

3. Sometimes, if the aim is prediction, it may be expensive (or, e.g. in some medical applications, risky or painfull) to observe some of the explanatory variables and it would be beneficial if it turned out to be unnecessary to measure some of the variables in future in order to still achieve an acceptable predictive accuracy. (Note, however, that in such a situation often not all variables are equally costly. Individual tests whether the most expensive variables are needed may then be more sensible than the application of general methods of variable selection.)

There are also arguments against variable selection:

1. In terms of predictive accuracy, as long as enough observations are available, it is usually much worse to leave out variables that are really important than to keep variables in the

model of which the true coefficients are (about) zero. The reason is that with enough observations the true zero coefficients will be estimated to be about zero anyway. There- fore, variable selection should not be done unless there is a real need for it (i.e. one of the reasons given above).

2. If some of the variables in the full model are highly correlated (i.e. they measure more or less the same thing), usually not all of these variables are needed, but the decision about precisely which variable should be kept in the model and which should be left out is quite arbitrary. Therefore, variable selection is often unstable and the chosen variables have to be interpreted with care concerning explanation and causal inference. Particularly, it cannot be taken for granted that a variable left out by variable selection is "really unimportant", because it may just be "represented" by another variable still in the model.

4.3.1 Criteria to Compare Models

Several criteria have been suggested in the literature that take into account the number of explanatory variables in the model. In the present course, only two of them are introduced explicitly. More can be found, e.g., in Hastie, Tibshirani, and Friedman (2001).

Akaike Information Criterion (AIC)

For quite general models fitted by maximum likelihood, the Akaike Information Criterion (AIC) is defined as:

$$AIC = -2\hat{\ell}(\text{model}) + 2p$$

where $\hat{\ell}(\text{model})$ is the maximum of the log-likelihood function, and p is the number of regression parameters in the model (including the intercept), i.e., $p = k+1$ for a model with k explanatory variables.

AIC is reported in R-software. Because a good model has a high likelihood and (hopefully) not many parameters, we look for models with the smallest AIC.

Under the assumptions of this chapter, $\hat{\ell}(\text{model})$ can be described by:

$$\hat{\ell}(\text{model}) = -\frac{n}{2}\log(\sigma^2) - \frac{RSS}{2\sigma^2} + \text{constant}.$$

Here, σ^2 is usually unknown and can be estimated by its Maximum Likelihood Estimator (MLE) $\hat{\sigma}^2_{\text{ML}} = \frac{RSS}{n}$. Hence, the AIC becomes:

$$AIC = n\log\left(\frac{RSS}{n}\right) + 2p.$$

The AIC is a so-called "penalised likelihood-method", because the term $2p$ can be interpreted as a penalty for too large models, correcting for the fact that the RSS is always decreased by increasing the model. For fixed k, the AIC just chooses the model with the smallest RSS.

Another criterion that is often used is the Bayesian Information Criterion (BIC)

$$\text{BIC} = -2\log(\ell) + p\ln(n)$$

where $\log()$ is the natural logarithm of the likelihood function evaluated at the parameter values that maximize it, p is the number of parameters in the model including the intercept, n is the number of observations. The term $p\log(n)$ acts as a penalty for complexity, particularly effective in larger samples, helping to avoid overfitting by penalizing models with more parameters.

Other functions of the RSS penalising large models have been suggested, such as Mallows C_p

$$C_p = \frac{RSS}{\hat{\sigma}^2} - n + 2p$$

and the so-called "adjusted R^2", which delivers a number between 0 and 1 like R^2 but is maximised if $\hat{\sigma}^2$ is minimised. This is, however, a very "soft" criterion which often does not reduce the number of variables enough (or even not at all).

In the following, you will explore various covariate selection techniques that have been proposed in the literature; those presented are just a small part. The important thing is that model selection should be done using appropriate criteria based on the likelihood function (e.g., AIC, BIC) and should never be done by comparing the p-value of the coefficients.

4.3.2 Best subset selection

In regression analysis, particularly when dealing with multiple explanatory variables, a crucial task is selecting the most effective model from a potentially vast number of possibilities. With p explanatory variables, you can construct 2^p different regression models, each representing a unique combination of variables.

Best Subset Selection: Best subset selection involves evaluating all 2^p possible models to identify the one that performs best according to certain criteria. We can breakdown the algorithm in the following steps

(a) **Evaluate Models for Each Subset Size**:

- For each possible number of explanatory variables, k, ranging from 1 to p, identify the best model.
- The "best" model for a given k is typically defined by the smallest Residual Sum of Squares (RSS). A smaller RSS indicates that the model fits the data more closely.
- Minimizing RSS is mathematically equivalent to minimizing the estimated variance of the errors ($\hat{\sigma}^2$), maximizing the coefficient of determination (R^2), and minimizing the p-value of the F-test. The F-test checks the null hypothesis that all regression coefficients are zero against the alternative that at least one is not.

(b) **Comparing Across Different k**:

- Adding more variables to a model will always decrease the RSS and increase R^2 because each additional variable can explain a part of the variance in the data, regardless of whether it is statistically significant.
- However, the challenge arises when comparing models with different numbers of variables (k). An increase in k usually means an improved fit (lower RSS) but doesn't necessarily indicate a better model due to potential overfitting or inclusion of irrelevant variables.

(c) **Statistical Challenges**:

- The model with $l > k$ variables does not necessarily include all the variables of the best model with k variables. This non-nesting of models complicates comparisons using standard t- and F-tests, which assume model nesting.

Advanced Model Comparison Techniques:

- Since comparing models of different sizes directly using RSS, R^2, or F-tests is problematic, other criteria such as Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) are used. These criteria help balance model fit with complexity, penalizing the addition of extraneous variables.

Considering Multiple Good Models:

- It's often advisable to identify several models that perform well rather than selecting a single "best" model. This approach acknowledges that data can support multiple plausible explanatory frameworks, especially in complex real-world scenarios.

This structured approach ensures that you not only find models that fit the data well but also guard against overfitting, thereby enhancing the robustness and interpretability of your results.

Algorithm 1 Best Subset Selection Algorithm for Implementation in a Programming Language (e.g., in R or Python).

Input: Set of p explanatory variables, dataset D
Output: Best model for each subset size
for $k = 1$ to p **do**
  Initialize BestModel$_k \leftarrow$ null
  Initialize BestScore$_k \leftarrow \infty$
  for all subsets S of size k of the explanatory variables **do**
    Fit model M_S using variables in subset S
    Calculate RSS or $\hat{\sigma}^2$ for model M_S
    if RSS of $M_S <$ BestScore$_k$ **then**
      BestModel$_k \leftarrow M_S$
      BestScore$_k \leftarrow$ RSS of M_S
    end if
  end for
  Print "Best model for k variables: ", BestModel$_k$
end for=0

4.3.3 Stepwise Methods

Stepwise methods are utilized to address the challenge of variable selection, particularly when the number of explanatory variables is large, making best subset selection computationally unfeasible. These methods streamline the process by significantly reducing the number of candidate models that need to be evaluated.

Backward Elimination

Backward elimination begins with a full model that includes all explanatory variables and systematically reduces the number of variables one at a time based on certain criteria:

Algorithm 2 Backward Elimination

0: **Input:** Dataset D, number of variables p
0: **Output:** Optimal model
0: Start with the full model using all p variables
0: Set $k = p$
0: **while** $k > 0$ **do**
0: Fit all k models, each excluding one of the variables from the current model
0: Select the model with the minimal Residual Sum of Squares (RSS)
0: Update the current model to this new model
0: $k = k - 1$
0: **end while**
0: **return** the final model =0

The total number of models evaluated is reduced to $\frac{p(p+1)}{2}$. After completing the backward sequence, selection criteria such as AIC, BIC, or cross-validation are typically used to determine the optimal model. Nested model comparison is feasible due to the sequential nature of the method, allowing the use of t-tests or F-tests.

Forward Selection

Forward selection starts with no variables in the model and adds one variable at a time, assessing each new model's performance:

Algorithm 3 Forward Selection

0: **Input:** Dataset D, number of variables p
0: **Output:** Optimal model
0: Begin with a model with no variables
0: Set $k = 0$
0: **while** $k < p$ **do**
0: Fit all $p - k$ models, each adding one new variable to the current model
0: Select the model with the lowest RSS
0: Update the current model to this new model
0: $k = k + 1$
0: **end while**
0: **return** the final model =0

This method also generates a nested sequence of models with the total number of models evaluated being $\frac{p(p+1)}{2}$. The same selection criteria used in backward elimination can be applied to choose the optimal model. However, less stringent significance levels might be considered to prevent the exclusion of informative variables prematurely.

Some General Remarks About Stepwise Methods

Stepwise methods generate heuristically reasonable sequences of models; however, they are not guaranteed to identify the best subset model. This is due to the fact that not all possible subsets of the covariates are evaluated. Consequently, best subset selection is superior as long as it remains computationally feasible.

Backward elimination and forward selection may not converge on the same sequence of models, nor necessarily identify the same "best" model. These methods are often quite unstable, meaning that minor changes in the data can result in significantly different model selections. This issue is somewhat less pronounced in best subset selection but still present.

Empirical evidence often shows that backward elimination tends to outperform forward selection in terms of prediction quality. However, if the number of observations n is very small relative to the number of covariates p (specifically, if $n < 2p$), the initial model in backward elimination becomes very unstable, making forward selection a preferable approach.

In particular fields such as genetics, where datasets may have $n < p$, backward elimination is not feasible (nor is fitting any regression model with more than $n - 2$ variables), but forward selection can still be effectively applied for sufficiently small values of k.

Implementation of Stepwise Methods in Statistical Software

Implementing stepwise selection methods can be efficiently done using popular statistical software such as R and Python. Below, we outline the functions and packages commonly used to apply backward elimination and forward selection.

Implementation in R

In R, the `step` function from the `stats` package is typically used for both backward elimination and forward selection. This function performs stepwise model selection by AIC.

- **Backward Elimination:** You can start with a full model and use the `direction="backward"` argument to perform backward elimination.

```
full.model <- lm(Y ~ ., data=dataset)
backward.model <- step(full.model, direction="backward")
```

- **Forward Selection:** Start with a model with no predictors, and use `direction="forward"` to add variables one by one.

```
null.model <- lm(Y ~ 1, data=dataset)
forward.model <- step(null.model,
direction="forward", scope=(~ ., data=dataset))
```

4.3.4 The Lasso

The Lasso (Least Absolute Shrinkage and Selection Operator) method is a regression technique used for variable selection and regularization, particularly effective in scenarios involving collinearity or a high number of predictors. Compared to other methods, Lasso offers greater stability and typically results in lower prediction errors. However, it introduces a moderate bias and may not perform optimally when true regression coefficients are either very close to zero or very large.

Standardization of Variables

To prepare for Lasso regression (similarly as we do for the Ridge), original variables (denoted as $z_1, \ldots, z_m$) must first be standardized. Each variable z_j is transformed into x_{ij} such that the new variables have zero mean and unit standard deviation, calculated using:

$$x_{ij} = \frac{z_{ij} - \overline{z}_j}{s_j}$$

where $\overline{z}_j$ and s_j represent the mean and standard deviation of the j-th variable, respectively. This standardization is crucial as it ensures the comparability of regression coefficients, reflecting their relative contributions to the model without implying their absolute importance due to potential dependencies among variables.

Lasso Estimation

The Lasso estimator $\hat{\beta}^L$ is derived by minimizing the sum of squared errors:

$$S(\beta) = \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

subject to the constraint that the sum of the absolute values of the coefficients does not exceed a predefined constant t:

$$\sum_{j=1}^{p} |\beta_j| \leq t$$

Notice the similarity with the ridge regression problem, in this case the ℓ_2 penalty $\sum_{j=1}^{p} \beta_j^2$ is replaced with the ℓ_1 penalty $\sum_{j=1}^{p} |\beta_j|$, where $|\cdot|$ denotes the absolute value. This latter containt makes the solution not linear, hence there is no closed form.

Computing the lasso solution is a quadratic programming problem, efficient algorithms are available for computing the entire path of solutions as λ is varied, with the same computational cost as for ridge regression. Because of the nature of the constraint, making t sufficiently small will cause some of the coefficients to be exactly zero. Thus, the lasso does a kind of continuous subset selection. If t is chosen larger than $t_0 = \sum_{j=1}^{p} |\hat{\beta}_j^{\text{OLS}}|$ (where $\hat{\beta}_j^{\text{OLS}}$ are the least squares estimates), then the lasso estimates are the $\hat{\beta}_j^L$'s. On the other hand, for $t = t_0/2$ say, then the least squares coefficients are shrunk by about 50% on average. The problem is that t should be adaptively chosen to minimize an estimate of expected prediction error.

Lasso is known as a shrinkage method because it reduces the absolute values of the coefficients. This reduction can significantly improve model accuracy by mitigating overfitting and biases associated with variable selection, common in methods like stepwise or best subset selection. However, the optimal degree of shrinkage depends on the true coefficient values and the error variance σ^2, posing challenges in setting a definitive guideline for t. To perform Lasso regression in R, one of the most widely used packages is `glmnet`. This package not only supports Lasso (ℓ_1 regularization) but also Elastic Net models, which combine ℓ_1 and ℓ_2 regularization techniques.

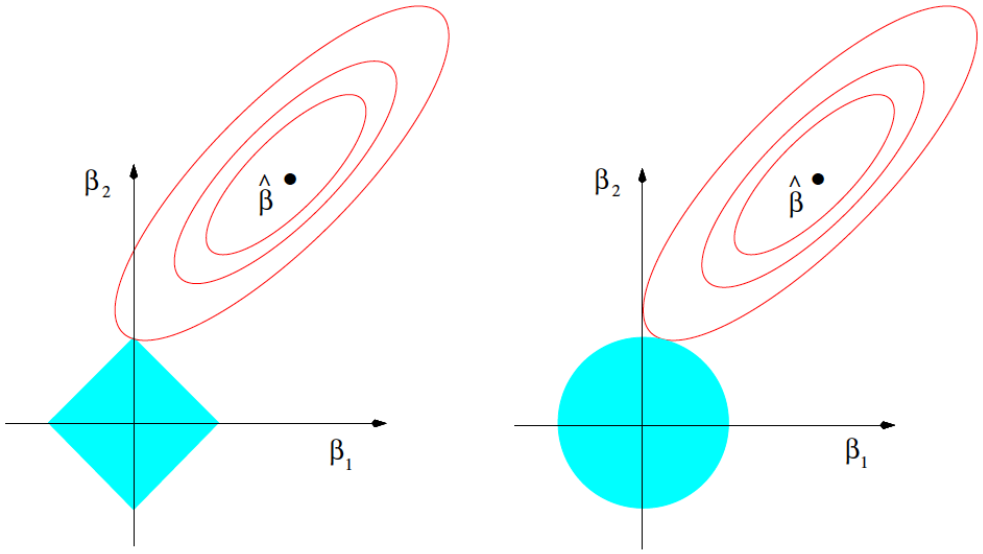


Figure 4.1: The estimation approaches for the lasso (left) and ridge regression (right) are depicted in the figures below. Shown are the contours of the error functions and the constraint functions. The solid blue areas represent the constraint regions for lasso and ridge regression, defined by $|\beta_1| + |\beta_2| \leq t$ and $\beta_1^2 + \beta_2^2 \leq t^2$, respectively. Meanwhile, the red ellipses illustrate the contours of the least squares error function. *source: Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction (Vol. 2, pp. 1-758). New York: springer.*

Chapter 5

Logistic Regression

Despite the simple linear model and its multivariate variant being very powerful tools for predicting qualitative variables (when Y assumes real values), they are less effective when our goal is to predict probabilities. This is because the linear model is ill-suited to this task due to the fact that $\mathbf{x}^\top \boldsymbol{\beta}$, once estimated, assumes real values. Therefore, we run the risk of predicting probabilities that are less than zero or greater than one, which is absolutely unacceptable, as it is well-known that probabilities must be between zero and one, see Figure **??** and its caption. For this reason, we need to introduce a new class of models, Generalized Linear Models. In this chapter we will present a specific type, the logistic model. Suppose that n responses $Y_1, \ldots, Y_N$ (in this section the sample size will be denoted with N, while with n we denote the number of trials of a Binomial random variable) are independent such that

$$Y_i \sim \text{Bin}(n_i, \pi_i)$$

with $\text{Bin}(n, p)$ we denote a Binomial Distribution. Where

$$g(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta}$$

link function $g(\cdot)$ relates the linear predictor $x_i^\top$ to the mean $\mu_i = E(Y_i) = n_i \pi_i$ for $i = 1, 2, \ldots, N$. This formulation includes the binary case in which $n_i = 1$ for all i. Note that the models below are given in terms of π_i. This is equivalent to expressing via μ_i as $\pi_i = \frac{\mu_i}{n_i}$.

The link function is the logit link, which gives the linear logistic model

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \text{logit}(\pi_i) = \mathbf{x}_i^\top \boldsymbol{\beta}$$

In all these cases, the expression for the probability of success, π_i, obtained by inverting the equation for the model, has been chosen to be a cumulative distribution function (cdf):

$$\pi_i = \frac{1}{1 + e^{-\eta_i}}, \quad \text{cdf of a logistic distribution.}$$

Here, $\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}$ denotes the linear predictor for any observation. Note that these equations can be used for prediction, after η_i has been estimated by $\hat{\eta}_i$, by plugging in $\hat{\eta}_i = \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}$.

The above choices ensure that $0 \leq \pi_i \leq 1$, as required. Also, the resulting link function has the property that $-\infty < g(\mu_i) < \infty$, with the consequence that there are no constraints on the unknown parameters in the linear predictor.

Indeed, what would happen if we wanted to predict π_i using a linear model of the same class introduced in the previous chapters? We would end up obtaining probabilities that are negative and greater than one.

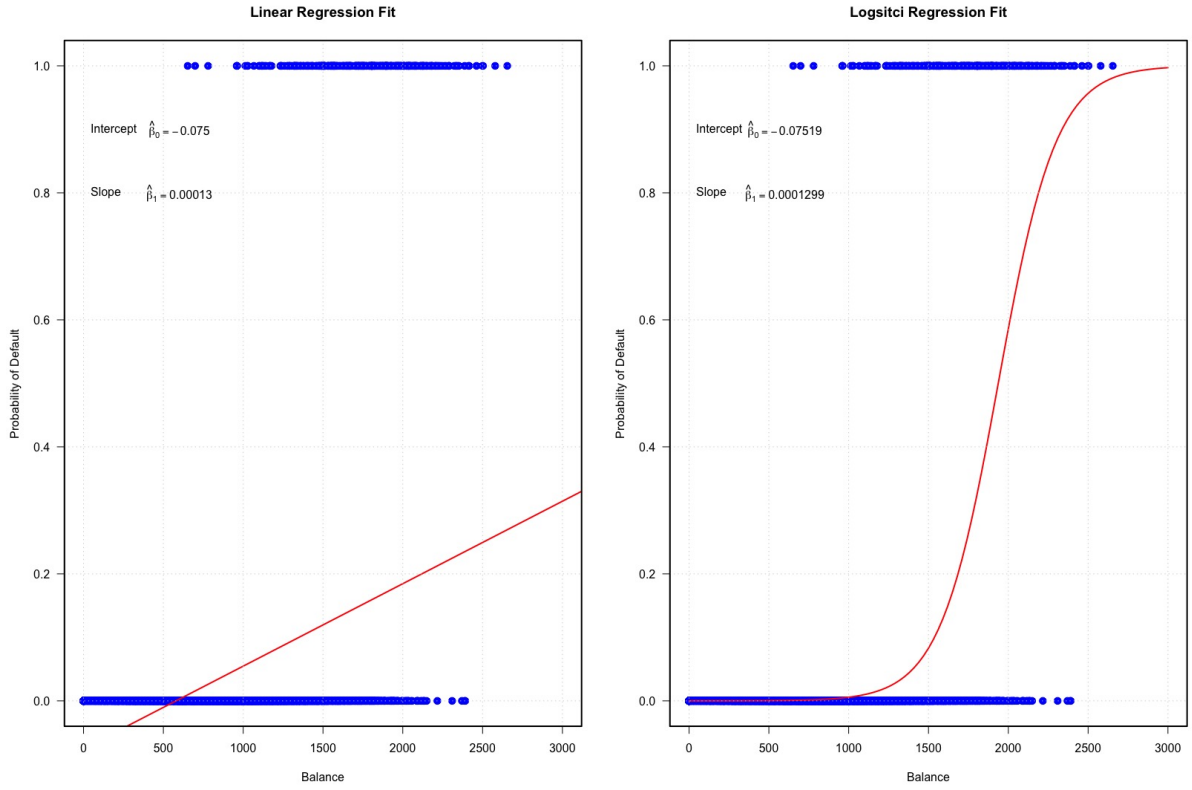


Figure 5.1: In this example, we wanted to predict the probability of default for some bank customers. On the right, we notice how the linear model fails to capture the probability. Not only does the fitted line seem implausible, but it also yields probabilities that are negative or greater than 1. On the left, however, we have a logistic model that correctly captures the probability levels for different values of the balance.

If π denotes the probability of a success, then the logit link function $\log\left(\frac{\pi}{1-\pi}\right)$ is the logarithm of the odds on a success, or 'log odds' for short, which we will denote by $\text{logit}(\pi)$. So the coefficient β_j of an explanatory variable x_j in a logistic regression model measures the rate of change of the log odds with x_j, holding constant the values of the other explanatory variables in the model.

In particular, suppose that x_1 is an indicator variable with just two levels, 0 and 1, as would be

the case if these values represented 'absence' and 'presence' of a factor. Then at $x_1 = 0$,

$$\text{logit}(\pi) = \beta_0 + \beta_2 x_2 + \ldots + \beta_p x_p,$$

and at $x_1 = 1$,

$$\text{logit}(\pi') = \beta_0 + \beta_1 + \beta_2 x_2 + \ldots + \beta_p x_p,$$

where the probability of success in the second case has been denoted by π'.

Subtracting these equations gives

$$\beta_1 = \text{logit}(\pi') - \text{logit}(\pi) = \log\left(\frac{\pi'}{1-\pi'}\right) - \log\left(\frac{\pi}{1-\pi}\right) = \log\left(\frac{\pi'/(1-\pi')}{\pi/(1-\pi)}\right),$$

which is the logarithm of the ratio of the odds on a success at the two values of x_1, or logarithm of the odds ratio, holding constant the values of the other explanatory variables in the model.

So, the odds on a success when $x_1 = 1$ is e^{β_1} times the odds on a success when $x_1 = 0$, holding constant the values of the other explanatory variables in the model.

Example 6. *To study the onset of cardiovascular diseases $Y = 1$ in relation to smoking habits (FU), the logit model is*

$$\pi = \mathbb{P}(Y = 1) = \frac{1}{1 + e^{-\beta_0 - \beta_1 FU}}$$

which means

$$\textit{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 \boldsymbol{FU}$$

So, if we obtain an estimated value of $\hat{\beta}_1 = 0.693$, this means that the transition from Smoking=NO to Smoking=YES results in a multiplicative increase in the log-odds of 0.693, which corresponds to an odds ratio of $e^{0.693}$. This implies that for smokers, the probability of becoming ill compared to not becoming ill is in a 2 to 1 ratio compared to non-smokers.

5.1 Estimation

The log-likelihood $\ell(\pi)$ for a binomial logistic regression model is given by:

$$\ell(\pi) = \sum_{i=1}^{N} \left[y_i \log(\pi_i) + (n_i - y_i) \log(1 - \pi_i) + \log \binom{n_i}{y_i} \right]$$

where $n_i = 1$ if Y is distributed according to a Bernoulli variable, by solving the estimation equations

$$\frac{\partial \ell(\pi)}{\partial \pi_i} = 0 \;\; i = 1, \ldots, N$$

For example, if you want to estimate the parameters of a model with a single explanatory variable $\text{logit}(\pi_i) = \beta_0 + \beta_1 x_i$ the log-likelihood becomes

$$\ell(\pi) = \sum_{i=1}^{N} \left[y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + n_i \log\left(1-\pi_i\right) + \log\binom{n_i}{y_i} \right]$$

it results in

$$\ell(\pi) = \sum_{i=1}^{N} \left[y_i(\beta_0 + \beta_1 x_i) - n_i \log\left(\beta_0 + \beta_1 x_i\right) + \log\binom{n_i}{y_i} \right]$$

therefore, it follows

$$\frac{\partial \ell(\pi)}{\partial \beta_0} = \sum_{i=1}^{N} \left[y_i - n_i \frac{e^{(\beta_0+\beta_1 x_i)}}{1+e^{(\beta_0+\beta_1 x_i)}} \right] = \sum_{i=1}^{N} \left[y_i - n_i \pi_i \right]$$

and

$$\frac{\partial \ell(\pi)}{\partial \beta_1} = \sum_{i=1}^{N} \left[y_i x_i - n_i \frac{e^{(\beta_0+\beta_1 x_i)} x_i}{1+e^{(\beta_0+\beta_1 x_i)}} \right] = \sum_{i=1}^{N} \left[y_i x_i - n_i \pi_i x_i \right]$$

Deriving the score function yields the information matrix and thus the asymptotic covariance matrix, from which the standard error estimates are derived:

$$se(\beta_0) = \left(\sum_{i=1}^{N} [n_i \pi_i (1-\pi_i)] \right)^{-1/2}$$

$$se(\beta_1) = \left(\sum_{i=1}^{N} [n_i x_i^2 \pi_i (1-\pi_i)] \right)^{-1/2}$$

5.2 Inference

i **Sampling Distribution of $\hat{\beta}$:**

$$\hat{\beta} \sim \mathcal{N}_{p+1}(\beta, \text{se}(\hat{\beta})).$$

ii **Deviance for Testing the Goodness-of-Fit of a Model:** The deviance for testing the goodness-of-fit of a particular model is given by

$$D = 2 \sum_{i=1}^{N} \left[y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) + (n_i - y_i) \log\left(\frac{n_i - y_i}{n_i - \hat{\mu}_i}\right) \right],$$

where the $\hat{\mu}_i$'s are the fitted μ_i's under the model (using the convention that $0 \log 0 = 0$). If the model is true, $D \sim \chi^2_{n-p+1}$, which provides a test statistic for a goodness-of-fit test.

iii **Test for $H_0 : \beta_j = 0$:**

$$\frac{\hat{\beta}_j}{\text{se}(\hat{\beta}_j)} \sim N(0,1) \text{ under } H_0.$$

This is a test for the omission of the j-th explanatory variable given the other explanatory variables in the model.

iv **Test for $H_0 : \nu$ of the Regression Parameters $\beta_1, \ldots, \beta_p$ are 0:** Here we are assuming that the linear predictor consists of a constant term and terms from p explanatory variables and H_0 tests the omission of ν explanatory variables where $\nu \leq p$. Let D_0 and D denote the deviances under H_0 and the maximal (full) models, respectively. The likelihood ratio test for H_0 is

$$D_0 - D \sim \chi^2_\nu \text{ under } H_0.$$

v An alternative test for assessing goodness-of-fit is the use of the Pearson chi-squared statistic, denoted by X^2. This test is based on the comparison of observed and fitted frequencies (the latter are usually referred to as expected frequencies in introductory texts, but are essentially estimates known as fitted values or, in this context, fitted frequencies):

Observation	1	2	...	N
Observed # successes	Y_1	Y_2	...	Y_N
Observed # failures	$n_1 - Y_1$	$n_2 - Y_2$	...	$n_N - Y_N$

Table 5.1: Table of Observed and Fitted Frequencies

A similar table of fitted values is used, where Y_i is replaced by $\hat{\mu}_i = n_i\hat{\pi}_i$.

Aside: Using common notation o for observed frequency and e for fitted (expected) frequency, D and X^2 have the forms:

$$D = 2\sum o \log\left(\frac{o}{e}\right) \quad \text{and} \quad X^2 = \sum \frac{(o-e)^2}{e}.$$

This form of the deviance D is often denoted by G^2.

The Pearson chi-squared statistic, after some algebraic manipulations, is given by:

$$X^2 = \sum_{i=1}^{N} \frac{(Y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i(1-\hat{\pi}_i)}$$

which, if the model is correctly fitted, has the same large sample distribution as the deviance D, i.e., χ^2_{N-p+1}. By using the Taylor Series expansion of $s\log(s/t)$ about $s = t$ up to the quadratic term, it can be shown that $D \approx X^2$.

The large sample distribution for D and X^2 is likely to be poor if any of the fitted values in the $2 \times N$ table are small.

Besides considering the residual deviance D, model adequacy should also be checked by appropriate plots (essentially the same that we have seen for the Linear model).

There are several forms of residuals in the binomial case.

- **Raw residuals:** $\hat{e}_i = Y_i - n_i\hat{\pi}_i$
  Interpretation: these show how well the raw data are fitted.

- **Pearson or chi-squared residuals:**

$$X_i = \frac{\hat{e}_i}{\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)}}$$

  so that the chi-squared statistic $X^2 = \sum_{i=1}^{N} X_i^2$.
  Interpretation: these standardise the raw residuals by the estimated standard deviation of Y_i, making them comparable in size.

- **Standardised Pearson residuals:**

$$r_{Pi} = \frac{X_i}{\sqrt{1-h_{ii}}}$$

  where h_{ii}, the leverage for the ith observation, is the ith diagonal element of the hat matrix, these are standardised so that their variance is 1. They are comparable in size as well but adjust for the location in x-space. For mathematical comparison these are better than the Pearson residuals, but the Pearson residuals are more naturally interpreted in terms of which points are "well fitted".

- **Deviance residual:**

$$d_i = \text{sign}(\hat{e}_i)\left(2\left[y_i\log\left(\frac{y_i}{\hat{\mu}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - \hat{\mu}_i}\right)\right]\right)^{1/2},$$

  so that the deviance $D = \sum_{i=1}^{N} d_i^2$. The term $\text{sign}(\hat{e}_i)$ gives d_i the same sign as $\hat{e}_i$.
  Interpretation: these formalise how strongly the observation contributes to the deviance, i.e., to the standard way of measuring the quality of the overall fit (or rather misfit; the higher, the worse). They show to what extent the observation indicates that the model is violated and rather a saturated model is needed.

- Standardised Deviance Residual:

$$r_{Di} = \frac{d_i}{\sqrt{1-h_{ii}}}$$

  Interpretation: This modification makes the d_i directly mathematically comparable by unifying their variance and adjusting for the location in x-space.

5.2.1 Penalized Logistic Regression

Consider the case of logistic regression where $Y_i \sim \text{Bernulli}(\pi)$, i.e. $\text{Binomial}(1, \pi)$, with

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} = f_{\beta_0,\beta_1,\ldots,\beta_p}(x)$$

the negative likelihood equals to

$$\rho(x, y) = -\ell(\pi) = \sum_{i=1}^{N} \left[-y_i f_{\beta_0,\beta_1,\ldots,\beta_p}(x_i) + \log\left(1 + \exp\left(f_{\beta_0,\beta_1,\ldots,\beta_p}(x_i)\right)\right)\right]$$

The ℓ_1-norm penalized Lasso estimator is the defined as

$$\hat{\beta}_0(\lambda), \hat{\beta}_1(\lambda), \ldots, \hat{\beta}_p(\lambda) = \underset{\beta_0,\beta_1,\ldots,\beta_p}{\arg\min} \left\{ n^{-1} \sum_{i=1}^{N} \rho(x_i, y_i) + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

As can be inferred, the ℓ_2 penalty can also be applied in this context resulting in

$$\hat{\beta}_0(\lambda), \hat{\beta}_1(\lambda), \ldots, \hat{\beta}_p(\lambda) = \underset{\beta_0,\beta_1,\ldots,\beta_p}{\arg\min} \left\{ n^{-1} \sum_{i=1}^{N} \rho(x_i, y_i) + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

as it can be easily demonstrated that the log-likelihood of the binomial model is a convex function. This convexity property arises because the model belongs to the exponential family. Therefore, applying ℓ_1 and ℓ_2 penalties to logistic regression does not pose significant computational problems. The effects of these penalties are essentially similar to those observed for the multiple linear regression model, as described in previous paragraphs.

The `glmnet` function in R supports various types of models specified by the `family` parameter. These include:

- `"gaussian"` for linear regression,
- `"binomial"` for logistic regression,
- `"poisson"` for Poisson regression,
- `"multinomial"` for multinomial logistic regression,
- `"cox"` for Cox proportional hazards regression,
- `"mgaussian"` for multiple Gaussian output regression.

For implementing penalized logistic regression, we use `family='binomial'`, which is suited for binary outcomes.

Chapter 6

Useful R commands

Reading Data

```
dta <- read.csv(file=file.choose())
# Choses directly the csv file by clicking


dta <- read.table(file=file.choose())
# Choses directly the txt file by clicking


dta <- read.csv("/danilo/desktop/mydataset.csv")
# Reading a csv file using a file path,
# I reccomend to set your working directory first
```

Manipulate dataframes

```
dta
# Shows the  data set called dta

head(mydata)
# Shows the first 6 rows (Default)
# The parameter n controls how many rows to show

tail(mydata)
# Shows the last 6 rows (Default)
# The parameter n controls how many rows to show
```

```
str(mydata)
# Shows structure of the dataset
# Variable name, type, size of the data frame etc..


names(mydata)
# Returns the variable names


ls()
# Shows a list of objects that are available
```

Descriptive Statistics

Assume that $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^n$ are two data vectors

```
mean(x)
# Computes the mean of  x
# If you are dealing with NAs, the option na.rm=T could be an option
# (Not exclusive of mean(), just check the R documentation)

median(x)
# Computes the median of  x

var(x)
# Computes the unbiased variance of  x

sd(x)
# Computes the unbiased standard deviation of  x

IQR(x)
# Computes the Inter Quartile Range (IQR) which is IQR= Q3-Q1 of  x, where
# Q1 is the first quartile and Q3 the third quartile

summary(x)
# Computes the 5-number summary and the mean of  x

cor(x,y)
```

```
# Computes the unbiased correlation coefficient between x and y

cor(mydata)
# Computes the unbiased correlation matrix
```

Basic Graphs

```
hist(x)
# Returns a histogram

boxplot(x)
# Returns a boxplot

boxplot(y~x)
# Returns a side-by-side boxplots
# ( one of the variable have to be factor)

plot(y~x)
# Returns a scatterplot of y versus x

plot(dta)
# Returns a scatterplot matrix of the data.frame/matrix called dta

abline(lm(y~x))
#adds regression line to a scatterplot
```

Liner Regression

```
##########################################
#                 Fitting
##########################################

model_fit=lm(y~x)
# Fit a regression model

summary(model_fit)
# Get results from fitting the regression model
# (resudual distribution, coefficients, t tests, p-values,
```

```
# F test, R^2, Adjusted R^2)

anova(model_fit)
# get the ANOVA table

plot(model_fit)
# Get four plots, including normal probability plot, of residuals

confint(model_fit)
# CIs for all parameters

predict.lm(model_fit, interval="confidence")
# Make prediction and give confidence interval for the mean response

predict.lm(model_fit, interval="prediction")
# Make prediction and give prediction interval for the mean response

newx=data.frame(X=4)
# Create a new data frame with one new x* value of 4

predict.lm(model_fit, newx, interval="confidence")
# Get a CI for the mean at the value x*

################################################
#                    Tests
################################################

bptest(model_fit)
# Get the Breusch-Pagan test
(lmtest package must be installed)

ad.test(resids)
# Get Anderson-Darling test for normality
(nortest package must be installed)

cvm.test(resids)
# Get Cramer-von Mises test for normaility
(nortest package must be installed)
```

```
lillie.test(resids)
# Get Lilliefors (Kolmogorov-Smirnov) test for normality
(nortest package must be installed)

pearson.test(resids)
# Get Pearson chi-square test for normaility
(nortest package must be installed)

sf.test(resids)
# Get Shapiro-Francia test for normaility
(nortest package must be installed)

boxcox(model_fit)
# Evaluate possible Box-Cox transformations
(MASS package must be installed)

################################################
#           Variable Selection
################################################

stepAIC()
# Chose best model using AIC
# "both" (for stepwise regression, both forward and backward selection);
# "backward" (for backward selection)
# "forward" (for forward selection).
(MASS package must be installed)

step()
# Choose a model by AIC in a Stepwise Algorithm
```

Chapter 7

Lab in R

In this chapter, all the laboratories covered in class are compiled. The following material does not replace but complements the study of the lecture notes. To successfully pass the exam, we recommend a thorough study of the following material, as well as completing all the exercises provided.

7.1 Lab 1 Exploratory Data Analysis

Setup

```
# Most of the data we are going to work with are taken from faraway
   package
library(faraway) # call the package to use its built-in functions and
   data
library(ggplot2) # an amazing package to create graphs
```

Load and Examine Data

```
data(pima)
head(pima) # just print the first 6 rows of the dataset
```

To ensure a comprehensive understanding of the data, it is crucial to generate numerical summaries such as means, quantiles, standard deviations (SDs), maximum and minimum values. This process is essential in determining the integrity of the data and identifying any potential outliers or inconsistencies. As a statistician or data scientist, exploring the data should be the first step in problem-solving.

The dataset under consideration is derived from a study conducted by The National Institute of Diabetes and Digestive and Kidney Disease, which involved 768 adult female Pima Indians residing near Phoenix. The variables within the dataset include:

Variable	Description
`pregnant`	Number of times pregnant
`glucose`	Concentration of plasma glucose at 2 hours in an oral glucose tolerance test
`diastolic`	Diastolic blood pressure in mmHg
`triceps`	Triceps skin fold thickness in mm
`insulin`	2-hour serum insulin in μU/ml
`bmi`	Body mass index (BMI), where weight is measured in kg and height in m^2
`diabetes`	Diabetes pedigree function
`age`	Age in years
`test`	Test about signs of diabetes, coded zero if negative and one if positive

Table 7.1: National Institute of Diabetes and Digestive and Kidney Disease: Data Description.

To initiate the exploration of this dataset, we can use the function `summary()`.

```
summary(pima)
```

There is something that doesn't make sense to you ? It is virtually impossible for a patient to have no blood pressure. Blood pressure can be zero or near-zero for certain medical conditions, such as shock or cardiac arrest, and can also vary depending on the position of the body. However, it is highly unlikely for a healthy individual to have a blood pressure of zero.

Regarding the dataset description, it is possible that the value of zero has been used to indicate missing data, as it is a common practice to represent missing values as zeros in some datasets. It is also possible that the researchers did not obtain the blood pressures for some patients due to certain limitations or errors in data collection. It is important to check the data documentation and consult with experts in the field to gain a better understanding of the data and potential issues.

At this point, it makes sense to denote the zero values as `NA`.

```
pima$diastolic[pima$diastolic==0] <- NA
pima$glucose[pima$glucose==0] <- NA
pima$triceps[pima$triceps==0] <- NA
pima$insulin[pima$insulin==0] <- NA
pima$bmi[pima$bmi==0] <- NA
```

The variable `test` is a categorical variable, also called a factor. Therefore, we need to be sure that `R` treats qualitative variables as factors. Sometimes (even professional statisticians) forget this and compute statistics such as 'average zip code'.

Formatting variables is not only important for summary statistics but also when we move on to modeling. Don't neglect the formatting phase, please.

```
str(pima$test)
```

In this case, `test` results in being an integer; this does not make sense, so it is better to format it as a factor.

```
pima$test <- factor(pima$test)
summary(pima$test)
```

Now that it's coded correctly, `summary(test)` makes more sense to all of us. We see that 500 cases were negative, and 268 were positive. A way to make this clearer is to use descriptive labels.

```
levels(pima$test) <- c('negative', 'positive')
summary(pima)
```

Now we are ready to explore a little bit further with some plots.

```
hist(pima$diastolic,
     xlab='Diastolic',
     main='')
```

```
plot(density(pima$diastolic, na.rm=TRUE),
     main='Distolic Kernel plot ')
```

If you are uncomfortable with the Kernel methods, this is an extraordinary resource to explore this topic further CLICK HERE. The kernel plot effectively avoids the blockiness that can be distracting in a histogram. However, it is important to ensure appropriate bin specifications for the histogram and bandwidths for the kernel density plot. To understand the effect of bandwidths, we can play with it.

```
plot(density(pima$diastolic, na.rm=TRUE, bw=1),
     main='This is a too wiggly plot ')
```

```
plot(density(pima$diastolic, na.rm=TRUE, bw=3),
     main='This looks better... more smooth ')
```

```
plot(density(pima$diastolic, na.rm=TRUE, bw=10),
     main='This is perfect ')
```

The higher the bandwidth, the smoother the density estimate will be. An in-depth discussion regarding kernel methods is out of the scope of this course. If you feel like you don't know enough about Kernels, please visit this CLICK HERE.

```
plot(diabetes ~ diastolic, data=pima)
```

```
plot(diabetes ~ test, data=pima)
```

An alternative to the base plots in R is the ggplot2 package. The essential elements of a plot made using this package are:

- **Data**: The data that is being visualized is passed to ggplot2 as a data frame.
- **Aesthetic mapping**: The mapping of variables in the data to visual properties of the plot, such as the x and y axis or color and shape of points.
- **Geometric objects**: The geometric objects that define the type of plot, such as points, lines, bars, or histograms.

More information about how to visualize data properly is discussed in MA304-7.

```
ggplot(data= pima,
       aes(x=diastolic))+
        geom_histogram()
```

```
ggplot(data= pima,
       aes(x=diastolic))+
        geom_density()+
        ggtitle('Diastolic kernel density')
```

```
ggplot(data=pima,
       aes(x=diastolic, y=diabetes, shape=test))+
        geom_point()
```

This is an example of a bivariate scatterplot (two dimensions) to which a third has been added. Can you think of ways to add dimension to this scatterplot? To add dimensions to a scatterplot in R, various techniques can be employed. Some of the most common approaches are:

- Adding color: You can add color to the points in the scatterplot to represent a third variable. For example, you can assign different colors to the points based on a categorical variable.
- Adding size: You can adjust the size of the points to indicate a numerical variable. For instance, you can make the points larger or smaller based on a variable's value.
- Adding shape: You can change the shape of the points to indicate a categorical variable. For example, you can use different shapes, such as circles, triangles, or squares, to represent different categories.

- Adding facets: You can create a grid of scatterplots, each representing a subset of the data based on one or more variables. This approach is useful for visualizing complex relationships in the data.

```
ggplot(data=pima,
       aes(x=diastolic, y=diabetes))+
        geom_point(size=1)+
        facet_grid(~ test)
```

Exercises

Please attempt the following exercises. Recall that to attempt these questions, you need to install the `faraway` package first.

Exercise 1

The dataset `teengamb` concerns a study of teenage gambling in Britain (`?teengamb`, for further details about the data). Make a numerical and graphical summary of the data, commenting on any features you find interesting. Limit the output you present to a quantity that a busy reader would find sufficient to get a basic understanding of the data.

Exercise 2

The dataset `uswages` is drawn as a sample from the Current Population Survey in 1988. Make a numerical and graphical summary of the data.

References

- Faraway, J. (2015). Linear Models with R Second Edition CHAPMAN & HALL/CRC Texts in Statistical Science.

7.2 Lab 2: Estimation

Setup

```r
# Most of the data we are going to work with are taken from faraway
   package
library(faraway) # call the package to use its built-in functions and
   data
library(ggplot2) # an amazing package to create graphs
```

Estimation

We can start by recalling some basics notions from lecture notes, a multiple linear model can be defines as:

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

where $\beta_i \;\; i = 0, \ldots, p$ are unknown parameters and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top$ can be assumed to follow a Gaussian distribution (or not) with $\mathbb{E}(\boldsymbol{\epsilon}) = 0$ and $\text{Var}(\boldsymbol{\epsilon}) = \sigma^2 I_n$, where I_n is an identity matrix (more details in the lecture notes).

In statistical modeling, specifically under linear model(s) domain, we are assuming a linear relationship between the dependent variable and one or more independent variables or predictors, also known as covariates. The term "linear" in linear model refers to the fact that the parameters of the model enter linearly, meaning that the effect of each predictor on the dependent variable is assumed to be a linear function of that predictor. However, it is important to note that the predictors themselves do not have to be linear. In fact, the linear model can be used with non-linear transformations of the predictors, as long as the relationship between the transformed predictors and the dependent variable is still linear. This flexibility makes the linear model a powerful tool for analyzing a wide range of data sets, from simple to complex, and for making predictions based on those data. For example:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 \log X_2 + \beta_3 X_1 X_2 + \epsilon.$$

Despite the presence of a logarithmic term and an interaction term, the given equation is still considered a linear model. This is because the model is linear in its parameters, which are the coefficients (β_0, β_1, β_2, β_3) that multiply each of the predictor variables. The dependent variable Y is a linear combination of the parameters and the predictor variables, where the coefficients are constant and do not depend on any function or transformation of the predictor variables. The

logarithmic term and interaction term do not violate the linearity assumption of the model. But what about this:

$$Y = \beta_0 + \beta_1 X_1^{\beta_2} + \epsilon.$$

The given equation is not a linear model because it is not linear in the parameters. Specifically, the exponent β_2 applied to X_1 means that the coefficient β_1 is not constant, but rather varies as a power function of X_1. This violates the linearity assumption of the model, which states that the coefficients must be constant and not depend on any function or transformation of the predictor variables.

Nonlinear models require more complex techniques for estimation and inference.

Now let's take a look at an example to understand how things work in the `R` language. The dataset we are going to examine today concerns the number of species found on the various Galapagos Islands.

```
library(faraway)
data(gala)
head(gala)
```

The `gala` dataset includes several variables that describe different aspects of the Galapagos Islands. These variables are:

Variable	Description
`Species`	Number of plant species found on each island
`Endemics`	Number of endemic species on each island
`Area`	Size of each island in square kilometers
`Elevation`	Highest elevation on each island in meters
`Nearest`	Distance to the nearest neighboring island in kilometers
`Scruz`	Distance to Santa Cruz island in kilometers
`Adjacent`	Area of the adjacent island in square kilometers

Table 7.2: Description of Variables in Galapagos Islands Dataset

To learn more about these variables, you can access the data description by running the command `?gala` in `R`.

```
str(gala)
```

Everything seems to be formatted correctly. We can fit a linear model in **R** using the `lm()` function. `lm()` is a function in `R` that fits linear regression models to data. In brief, you need two ingredients:

1. You need to create a formula that specifies the relationship between the response variable (the variable you want to predict) and one or more predictor variables. The formula takes the form

```
response_variable ~ predictor_variable_1 + predictor_variable_2
```

For example, if you have a dataset with a response variable called `y` and two predictor variables called `x1` and `x2`, the formula might be

```
y ~ x1 + x2
```

2. Once you have the formula, you can use the `lm()` function to fit the linear regression model. The syntax for `lm()` is `lm(formula, data)`. The first argument is the formula, and the second argument is the name of the data frame that contains your variables.

```
lmod <- lm(Species ~ Area + Elevation + Nearest + Scruz + Adjacent,
   data = gala) # fit the model
summary(lmod)
```

From this output, we can extract the regression quantities we need:

R Function	Description
`residuals()`	Residuals (e_i): The differences between observed values (y_i) and predicted values ($\hat{y}_i$)
`fitted()`	Fitted Values ($\hat{y}_i$): The predicted values for each observation
`df.residual()`	Degrees of Freedom (DF): The degrees of freedom associated with the residuals
`deviance()`	Deviance (RSS): The residual sum of squares, a measure of model fit
`coef()`	Coefficients ($\hat{\beta}_i$): The estimated coefficients of the regression model

Table 7.3: Description of Quantities Returned by R Functions

```
residuals(lmod)
fitted(lmod)
# Please make sure that (species - fitted values) gives the residuals
df.residual(lmod)
# degrees of freedom are 24; does it make sense to you?
deviance(lmod)
coef(lmod)
```

You can even extract other quantities by examining the model object and its summary:

```
names(lmod)
```

```
lmodsum <- summary(lmod)
names(lmodsum)
```

Now we can check the assumptions:

```
plot(lmod$residuals)
```

```
qqnorm(lmod$residuals)
```

```
plot(lmod, 1:2)
```

The `plot()` function in `R` can be used to generate four diagnostic plots for linear regression models:

1. Residuals vs Fitted Values Plot: This plot shows the residuals (the differences between the observed values and the predicted values) on the y-axis and the fitted values (the predicted values) on the x-axis. The plot is used to check for non-linear relationships between the predictor variables and the response variable. If there is a clear pattern in the residuals, such as a curve or a funnel shape, it may indicate a non-linear relationship.

2. Normal Q-Q Plot: This plot shows the residuals on the y-axis and their expected values under normality on the x-axis. The plot is used to check for violations of the assumption that the residuals are normally distributed. If the residuals are normally distributed, they will follow a straight line on the plot.

3. Scale-Location Plot: This plot shows the square root of the absolute values of the standardized residuals on the y-axis and the fitted values on the x-axis. The plot is used to check for violations of the assumption that the variance of the residuals is constant across the range of the predictor variables. If the variance is constant, the points on the plot will be evenly spread out around a horizontal line.

4. Residuals vs Leverage Plot: This plot shows the leverage (a measure of how much influence an observation has on the model) on the x-axis and the standardized residuals on the y-axis. The plot is used to check for influential outliers. If an observation has high leverage and a large standardized residual, it may be an influential outlier.

In linear regression, it is also useful to have some measure of goodness of fit. One common choice is the R^2. We recall here its formulation:

$$R^2 = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (y_i - \bar{y})^2}$$

R^2 is a statistical measure that represents the proportion of the variance in the dependent variable that is predictable from the independent variables in a regression model. It is commonly used as a goodness-of-fit measure for linear regression models.

The R^2 value ranges from 0 to 1, where 0 indicates that the model explains none of the variability in the dependent variable, and 1 indicates that the model explains all of the variability. However, it is rare to see an R^2 value of 1 in practice, as it would mean that the model perfectly predicts the dependent variable with no error.

In general, a higher R^2 value indicates a better fit of the model to the data. However, the interpretation of what constitutes a "good" R^2 value depends on the specific context and the goals of the analysis.

For example, in some fields, such as physics or engineering, models with R^2 values of 0.9 or higher may be considered good. In contrast, in social sciences or economics, where human behavior is much more complex and difficult to predict, an R^2 value of 0.2 to 0.3 may be considered a good fit.

It's important to keep in mind that R^2 should not be the only criterion for evaluating a model's performance, and it should be used in conjunction with other metrics such as residual plots, cross-validation, and statistical significance tests to ensure that the model is valid and reliable for the specific analysis at hand.

An alternative measure of fit is $\hat{\sigma}^2$. The advantage is that it is measured in the units of the response Y. The regression summary returns both values, and it is worth paying attention to both of them. It is good practice to explore your data before fitting any model you have in mind!

Diagnostics

Now we have estimated our linear model, it is time to check the assumptions seen in the lecture notes. It is a good practice to check regression diagnostics before using the model for any kind of inference purposes you have in mind (e.g., Tests, Confidence Intervals, Predictions).

We can divide our problems into two main categories:

1. We have assumed that $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_n)$, with this assumption we are not only telling that the error terms are all centered at zero and have constant variance, we are also saying that they are uncorrelated (Do you see why?).

2. We are assuming that the structural part of the model $\mathbb{E}(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}$ is correct and it is the same for all the observations; there is no structural break in the data.

Let's start by checking the independency, constant variance, and normality assumptions of the errors $\boldsymbol{\epsilon}$. As has been already discussed, the error terms are not observable, but we can estimate them using the residuals $\boldsymbol{e}$.

Merely examining the residuals by plotting them does not suffice to verify the assumption of constant variance, as other factors may be involved. To ensure accuracy, additional factors must be checked. One of the most crucial diagnostic plots is the plot of residuals versus predicted values. If all assumptions are valid, a symmetrical, constant variation in the vertical direction should be evident. This plot not only identifies heteroskedasticity but also detects non-linearity.

```
res1 <- rnorm(100)
res2 <- rnorm(100, sd=1:50)
res3.1 <- rnorm(33, mean=-8, sd=1)
res3.2 <- rnorm(33, mean=0, sd=1)
```

```
res3.3 <- rnorm(33, mean=-4, sd=1)
res3 <- c(res3.1, res3.2, res3.3)
par(mfrow=c(1,3))
{
  plot(res1, xlab='Fitted', ylab='Residuals', main='No problem')
  abline(h=0)
  plot(res2, xlab='Fitted', ylab='Residual', main='Heteroskedasticity')
  abline(h=0)
  plot(res3, xlab='Fitted', ylab='Residuals', main='Non-linearity')
  abline(h=0)
}
```

Creating a plot of residuals against a predictor variable that is not included in the model, such as $(e, \mathbf{x}_i)$, can also be helpful. If any patterns or structures are apparent in this plot, it may suggest that this predictor should be incorporated into the model.

It is often challenging to check the residuals without prior knowledge or experience with data, so we can generate some artificial plots just to get an idea of what a good (bad) model(s) looks like.

```
par(mfrow=c(1,3))
n <- 200
for (i in 1:3) {
  x <- runif(n)
  plot(x, rnorm(x), main='Constant Variance')
}
for (i in 1:3) {
  x <- runif(n)
  plot(x, x * rnorm(n), main='Non-Constant Variance')
}
for (i in 1:3) {
  x <- runif(n)
  plot(x, cos(x * pi) + rnorm(n, sd=1), main='Nonlinearity')
}
```

Since tests and confidence intervals rely heavily on the normality assumption, it is advisable to confirm its validity. The normality of residuals can be evaluated by examining a Q-Q plot, which compares the quantiles of the residuals with those of a normal distribution. Formally, a Q-Q plot, or quantile-quantile plot, is a graphical method used to compare the distribution of a sample to a theoretical distribution, such as a normal distribution. The plot displays the sample quantiles on the vertical axis and the theoretical quantiles on the horizontal axis. If the sample and theoretical distributions are identical, the points on the Q-Q plot will fall along a straight line.

Departures from a straight line indicate deviations from the theoretical distribution. Q-Q plots can be used to assess the normality of a distribution or to compare the distribution of a sample to other distributions, such as a uniform or exponential distribution.

```
data(savings)
lmod <- lm(sr ~ pop15 + pop75 + dpi + ddpi, savings)
{qqnorm(residuals(lmod))
qqline(residuals(lmod))}
```

Normal residuals should follow the 45-degree line approximately. Just to get an idea, we can simulate some non-normal distributions and see what a Q-Q plot looks like:

```
par(mfrow=c(1,3))
n <- 200

for (i in 1:3) {
  x <- rnorm(n)
  qqnorm(x)
  qqline(x)
}  # Normal distribution

for (i in 1:3) {
  x <- exp(rnorm(n))
  qqnorm(x)
  qqline(x)
}  # Lognormal distribution

for (i in 1:3) {
  x <- rcauchy(n)
  qqnorm(x)
  qqline(x)
}  # Cauchy distribution

for (i in 1:3) {
  x <- runif(n)
  qqnorm(x)
  qqline(x)
}  # Uniform distribution
```

Addressing non-normality is not a one-size-fits-all approach; the appropriate solution will depend on the nature of the problem encountered. In the case of skewed errors, transforming the response variable may be sufficient to resolve the issue. If non-normality persists despite a trans-

formation, an alternative approach may be to acknowledge the non-normality and make inferences based on a different distribution, or alternatively, employ resampling methods.

If you would like to test if the residuals follow a normal distribution or not, you can use the Shapiro-Wilk test.

```
shapiro.test(residuals(lmod))
```

Under the null, we have that the residuals are normally distributed. Since the p-value is large, we do not reject the null.

Exercises

Please attempt the following exercises. Recall that to attempt these questions, you need to install the `faraway` package first.

Exercise 1

1. Try to obtain the vector of coefficient estimates by using the formula in the lecture notes: $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y}$.

Exercise 2

Try to obtain $\hat{\sigma}^2$ from `lm()` output and compare it to the result obtained using the formula from the lecture notes.

Exercise 3

The dataset `teengamb` concerns a study of teenage gambling in Britain. Fit a regression model with the expenditure on gambling as the response and the sex, status, income, and verbal score as predictors.

- a) Try to replicate the output of the `summary()` function by computing all the quantities using the formulas seen in the lecture notes.
- b) What percentage of variation in the response is explained by these predictors?
- c) Which observation has the largest (positive) residual? Give the case number.
- d) Compute the mean and the median of the residuals. Are they symmetric and centered around zero?
- e) Compute the correlation of the residuals with the fitted values.

- f) Compute the correlation of the residuals with the income.
- g) For all other predictors held constant, what would be the difference in predicted expenditure on gambling for a male compared to a female?

Exercise 4

In this question, we investigate the relative merits of methods for computing the coefficients. Generate some artificial data:

```
x <- 1:20
y <- x + rnorm(20)
```

Fit a polynomial in `x` for predicting `y`. Compute $\hat{\boldsymbol{\beta}}$ in two ways: by using `lm()` and by using direct calculation described in the lecture notes. At what degree of polynomial does the direct calculation method fail? (Hint: you need to use the `I()` function when fitting the polynomial, i.e., `lm(y ~ x + I(x^2))`.

References

- Faraway, J. (2015). Linear Models with R Second Edition CHAPMAN & HALL/CRC Texts in Statistical Science.

7.3 Lab 3:Inference

Setup

```
# Most of the data we are going to work with are taken from faraway
   package
library(faraway) # call the package to use its built-in functions and
   data
library(ggplot2) # an amazing package to create graphs
```

Inference

With our model estimation complete, we can now move on to making inferences. It's important to note that to estimate the intercept β_0 and the parameters β_i for $i = 1, \ldots, p$, no particular assumptions are required, as OLS is simply an algebraic tool. However, if we wish to compute confidence intervals (CI) or perform hypothesis tests, we must make the assumption that the error terms are normally distributed. This is an important consideration when drawing conclusions from our model and ensuring the reliability of our results.

Tests about Multiple Parameters

Given several predictors, you might want to test whether all are needed. Consider a larger model, Ω of dimension p, and a smaller model, ω of dimension q, which consists of a subset of the predictors that are in Ω. We can obtain the following F statistic:

```
F = (RSS_omega - RSS_Omega / (p - q)) / (RSS_Omega / (n - p)) ~ F_{p-q,
   n-p}
```

Details about the derivation of this statistic are out of the scope of this course. However, F is obtained by a Likelihood ratio test, therefore by assuming that the error terms are Normally distributed. Hence, this test and most of the tests we are going to see in this section cannot be used if there is a strong violation of the Normality assumption.

In R, many tasks can be simplified. To illustrate this point, we will use the Galapagos Islands dataset to fit the same model as before, with the number of species as the response variable and the geographic variables as predictors.

```
lmod <- lm(Species ~ Area + Elevation + Nearest + Scruz + Adjacent,
   data=gala)
```

We then fit the model with only the intercept. The function that we are going to use is `anova()`; this will do the job for us!

```
nullmodel <- lm(Species ~ 1, data=gala)
anova(nullmodel, lmod)
```

We can see directly the result of the test, where the null here is

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$$

$$H_1 : \text{at least one is different from zero,}$$

we have a p-value of $6.838e-07$, we reject the null. This can be verified using the formula derived above

```
rss0 <- deviance(nullmodel)
rss <- deviance(lmod)
df0 <- df.residual(nullmodel)
df <- df.residual(lmod)

(fstatistic <- ((rss0 - rss) / (df0 - df)) / (rss / df))
```

we can see that this result is consistent with the one obtained using `anova()`.

```
1 - pf(fstatistic, df0 - df, df)
```

the p-value is consistent as well.

Testing one predictor

Lets suppose to test

$$H_0 : \beta_i = 0 \text{ vs } H_1 : \beta_i \neq 0.$$

Let Ω be the model with all the predictors of interest which has p parameters and let ω be the model with all the same predictors except predictor i (the one we are testing). Let's test whether `Area` can be dropped from the full model by testing the hypothesis that the corresponding parameter is zero.

```
lmod <- lm(Species ~ Area + Elevation + Nearest + Scruz + Adjacent,
   data=gala)
lmods <- lm(Species ~ Elevation + Nearest + Scruz + Adjacent, data=gala
   )
anova(lmods, lmod)
```

the p-value of 0.3 tells that we cannot reject the null. An alternative (equivalent) approach is to use t-statistic for testing the hypothesis:

$$t_i = \frac{\hat{\beta}_i}{se(\hat{\beta}_i)},$$

this test statistic can be proved to be distributed as a t-student with $n - (p + 1)$ degrees of freedom

Testing a pair of predictors

In statistical terms, if we want to investigate whether the area of either the current island or the adjacent island has any relationship with the response, we can phrase the problem as follows:

$$H_0 : \beta_{Area} = \beta_{Adjacent} = 0,$$

also this test can be carried out with `anova()`

```
lmods <- lm(Species ~ Elevation + Nearest + Scruz , data=gala)
anova(lmod, lmods)
```

The null is rejected as the p-value is too small.

Testing a subspace

Lets suppose that we want to test

$$H_0 : \beta_{Area} = \beta_{Adjacent},$$

we can test this hypothesis again with `anova()`

```
lmods <- lm(Species ~ I(Area + Adjacent) + Elevation + Nearest + Scruz
   , data=gala)
anova(lmods, lmod)
```

The function `I()` ensures that the argument is evaluated rather than interpreted as part of the model formula. The p-value suggests that the null can be rejected.

Yet, assume we want to test

$$H_0 : \beta_{Elevation} = 0.5.$$

This can be set in R by using an `offset`

```
lmods <- lm(Species ~ Area + offset(0.5 * Elevation) + Nearest + Scruz
   + Adjacent, data=gala)
```

```
anova(lmod, lmods)
```

we see that the p-value is small, so the null has to be rejected.

Confidence Intervals

In our domain, Confidence Intervals (CIs) are an incredibly powerful tool that allow us to quantify the uncertainty in our estimates of β. They provide us with a range of plausible values for the true value of the parameter, which can help us make informed decisions and draw accurate conclusions from our data.

$$\hat{\beta}_i \pm t_{n-(p+1),\alpha/2} se(\hat{\beta}),$$

```
lmod <- lm(Species ~ Area + Elevation + Nearest + Scruz + Adjacent,
   data=gala)
summary(lmod)
```

we can construct manually 95% CIs for β_{Area} for which we need the 2.5% and 97.5% percentiles of the t-distribution with $30 - 6 = 24$ degrees of freedom.

```
qt(0.975, 24)
```

```
-0.02394 + c(-1, 1) * qt(0.975, 24) * 0.02242
```

we can note that the CI contains zero, this indicates that the null hypothesis $H_0 : \beta_{Area} = 0$ would not be rejected at the 5% level.

A convenient way to obtain all the univariate intervals is

```
confint(lmod)
```

Exercises

Please attempt the following exercises. Recall that to attempt these questions, you need to install the `faraway` package first.

Exercise 1

Using the `prostate` data, fit a model with `lpsa` as the response and the other variables as predictors.

1. Compute 90 and 95% CIs for the parameter associated with `age`.

2. Using just these intervals, what could we have deduced about the p-value for `age` in the regression summary?

Exercise 2

Using the `sat` data

1. Fit a model `total` SAT score as the response and `expand`, `ratio`, and `salary` as predictors. Test the hypothesis that $\beta_{salary} = 0$. Test that $\beta_{salary} = \beta_{ratio} = \beta_{expand} = 0$. Do any of these predictors have an effect on the response?

2. Now add `takers` to the model. Test the hypothesis that $\beta_{takers} = 0$. Compare this model to the previous one using an F-test. Demonstrate that the F-test and the t-test here are equivalent.

Exercise 3

Find a formula relating R^2 and the F-test for the regression (please try to attempt this alone, not using external help or AI!).

References

- Faraway, J. (2015). Linear Models with R Second Edition CHAPMAN & HALL/CRC Texts in Statistical Science.

7.4 Lab 4: Prediction & Transformation

Predictions

Setup

```
# Most of the data we are going to work with are taken from faraway
   package
library(faraway) # call the package to use its built-in functions and
   data
library(ggplot2) # an amazing package to create graphs
```

Up until now, we've covered the process of estimating a linear model and making inferences based on that model. In this lab, our focus will shift towards the topic of prediction. The problem at hand can be summarized as follows: given a new set of predictors x_0, what is the predicted response? It can be proved to be:

$$\hat{y}_0 = x_0^T \hat{\beta}.$$

Here, $\hat{\beta}$ is estimated without taking into account the new information provided by x_0, hence we can express $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$. However, for this scenario, a point estimate is not sufficient to make reliable predictions. We need to assess the uncertainty associated with this prediction. For instance, if we want to predict the high water mark of a river, we may need to construct barriers that are high enough to withstand floods that could potentially exceed the predicted maximum. Financial projections are similarly not very useful without a realistic estimate of the associated uncertainty.

Just before introducing you to confidence intervals, we need to define what we mean by the predicted mean response and a prediction of a future observation. Let's suppose we have built a regression model that predicts the rental prices of houses in a given area based on predictors such as the number of bedrooms and proximity to a major highway. We have two types of predictions that can be made:

1. Suppose that a specific house comes to the market were its characteristics are in the vector x_0. Its rental price will be $Y_0 = x_0^T\boldsymbol{\beta} + \epsilon_0$, without loss of generality we can still assume that $\epsilon_0 \sim \mathcal{N}(0, \sigma^2)$. It is easy to prove that $E[\hat{Y}_0 - Y_0] = 0$ and $Var[\hat{Y}_0 - Y_0] = \sigma^2[x_0^T(\mathbf{X}^T\mathbf{X})^{-1}x_0 + 1]$ (please attempt to do this). In assessing the variance of the prediction $x_0^T\hat{\boldsymbol{\beta}}$ we need to take into account the variance of ϵ_0.

2. On the other hand, suppose our manager asks the question, "What would a house with characteristics x_0 rent for on average?" This average selling price is $x_0^T\boldsymbol{\beta}$ and is predicted

by $x_0^T\hat{\boldsymbol{\beta}}$. In this case, it makes sense to take into account only the variance of $\hat{\boldsymbol{\beta}}$. The recall in the first case we are looking for a **prediction of a future value**, while in the second case **prediction of the mean response**.

For (1) a $100(1-\alpha)\%$ CI for a single future response is

$$\hat{y}_0 \pm t_{n-(p+1),\alpha/2}\hat{\sigma}\sqrt{1 + x_0^T(\mathbf{X}^T\mathbf{X})^{-1}x_0}.$$

So far, CIs have been for parameters; under our Frequentist framework, we are assuming that a parameter θ is a fixed quantity, they are not random. However, a future observation by definition is random. For this reason, the interval above is called **prediction interval**. We say that there is a 95% chance that the future value falls within this interval. This would be incorrect and may be outrageous for some statisticians (including me) to say for a parameter. **Please take care about what you write in your lab exam**.

For (2) a $100(1-\alpha)\%$ CI is

$$\hat{y}_0 \pm t_{n-(p+1),\alpha/2}\hat{\sigma}\sqrt{x_0^T(\mathbf{X}^T\mathbf{X})^{-1}x_0}.$$

The CI for (2) is usually much narrower than the prediction interval. It is tempting to use this version to make intervals for predicted values; please do not do that. This is a serious mistake, especially if done for a report in a firm.

Let's see how to compute these quantities in R:

```
library(faraway)
data(fat)
lmod <- lm(brozek ~ age + weight + height + neck + chest + abdom + hip
   + thigh + knee +  ankle +  biceps + forearm + wrist, data= fat)
```

Measuring body fat is not an easy task. One method requires submerging the body underwater in a tank and measuring the increase in the water level. As many people would prefer not to be submerged underwater, researchers recorded age, weight, height, and 10 body circumference measurements for 252 with the aim to build a model able to predict body fat.

In the model just fitted, we use `brozek` as the response (Brozek's equation estimates percent body fat from density). Let's consider a typical man, this is exemplified by the median value of all the predictors:

```
x <- model.matrix(lmod) # this extracts the design matrix
x0 <- apply(x, 2, median)
x0
```

We obtained a vector with the same columns as the design but with just one row. We can obtain $\hat{y}_0$ using `predict`:

```
(y0 <- predict(lmod, new=data.frame(t(x0))))
```

Please take care as the `predict` function requires the new value argument being in the form of a data frame with the same format as the data used to fit the model.

Now if we want a 95% CI we have just decided whether we are predicting the body fat for one particular man or the mean body fat for all men with these same characteristics:

```
predict(lmod, new=data.frame(t(x0)), interval='prediction')
```

```
predict(lmod, new=data.frame(t(x0)), interval='confidence')
```

Transformation

Transformations can be an incredible tool to use to improve the fit and correct some violations (e.g., heteroscedasticity can usually be corrected by applying a $\log()$ transformation on the Y values). Another option is to add additional predictors that are functions of the existing predictors like quadratic or cross-product terms.

Suppose that we are considering the following model:

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

In the original scale of the response, this model becomes (by applying $\exp()$ to both sides):

$$Y = \exp(\beta_0 + \beta_1 X) \times \exp(\epsilon),$$

Under this framework, the model enters in a multiplicative way. Take-home message: using $\log()$ can save your model and your marks during the lab exam.

In practice, we may not know how the error enters the model. The best approach is to try different transformations to get the structural form of the model right and worry about the error component later. Non-linearity can often be detected by exploring your data through plotting (y vs x_i). Once a transformation is applied and the model is fitted, we can check the residuals to see whether they satisfy the conditions required for linear regression. When transforming, you have to take care when interpreting your data (more details in the lecture notes).

When you use the log on the independent variable, the regression coefficients can be interpreted as:

$$\log(\hat{y}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p,$$

$$\hat{y} = e^{\hat{\beta}_0} e^{\hat{\beta}_1 x_1} \ldots e^{\hat{\beta}_p x_p}.$$

An increase of one unit in x_1 would multiply $\hat{y}$ by $e^{\hat{\beta}_1}$. Thus, when a log scale is used, the regression coefficients can be interpreted in a multiplicative rather than an additive manner.

We can recall that $\log(1+x) \approx x$ (this is obtained by a Taylor expansion around zero). So for example, suppose to have $\hat{\beta}_1 = 0.09$; then an increase of one in x_1 would lead to about a 0.09 increase in $\log y$, which is a 9% increase in y.

Box Cox Transformation

A useful tool you can employ to get which transformation to apply is the Box-Cox method. This method is designed for strictly positive responses and chooses the transformation to find the best fit to the data. This method transforms the response $y \to g_\lambda(y)$ where the family of transformations is indexed by λ:

$$g_\lambda(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log(y) & \lambda = 0, \end{cases}$$

For fixed $y > 0$, $g_\lambda(y)$ is continuous in λ. The idea is to choose λ using the maximum likelihood method. Assuming the normality of the errors, the likelihood function is like this:

$$L(\lambda) = -\frac{n}{2}\log(RSS_\lambda/n) + (\lambda - 1)\sum_i \log(y_i),$$

where RSS_λ is the residual sum of squares when $g_\lambda(y)$ is the response. Transforming the response can make our model hard to interpret, so we want to be sure to do it only if necessary. One way to check this is to form a confidence interval of λ. A $100(1-\alpha)\%$ CI is:

$$\{\lambda : L(\lambda) > L(\hat{\lambda}) - 0.5\chi^2_{1,(1-\alpha)}\}.$$

For each value of λ, a transformation is proposed. Some are represented in the table below:

Lambda	Transformation
-3	Y^{-3}
-2	Y^{-2}
-1	Y^{-1}
-0.5	$Y^{-0.5}$
0	$\log(Y)$
1	Y
2	Y^2
3	Y^3

Let's move into R to see how things work. To use the Box-Cox transformation, we need to install and load the `MASS` library (I suggest you explore the documentation of `MASS` as it is an

interesting package with tons of statistical tools implemented).

```
library(MASS)
library(faraway)
data(savings)
lmod <- lm(sr ~ pop15 + pop75 + dpi + ddpi, savings)

boxcox(lmod, plotit = T, lambda = seq(0.5, 1.5, by = 0.1))
```

We can see from the plot that there seems to be no good reason to transform our data.

Let's consider the Galapagos Islands:

```
data(gala)

lmod <- lm(Species ~ Area + Elevation + Nearest + Scruz + Adjacent,
   gala)
summary(lmod)
plot(lmod)

boxcox(lmod, lambda = seq(-0.25, 0.75, by = 0.05), plotit = T)
```

A possible transformation is the square root, as this falls just within the confidence intervals.

```
lmod <- lm(sqrt(Species) ~ Area + Elevation + Nearest + Scruz +
   Adjacent, gala)
plot(lmod)
```

```
summary(lmod)
```

The residuals in the fitting with $\sqrt{Y}$ seem to be approximately normal (see the min, max, Q_1, and Q_2), and the residual standard error decreased.

Just some words of warning about Box-Cox transformation:

1. The Box-Cox is not robust against outliers, so if you get a value of $\hat{\lambda} = 5$, the reason could be that you are working with data affected by outliers. (Do you see why it's so important to explore your data first?)

2. If some $y_i < 0$, it is customary (not an elegant solution) to add a constant to all y's. Note that the constant to add should be small (e.g., 0.05, 0.10).

3. There is doubt about whether the estimation of λ counts as an extra parameter to be considered in the degrees of freedom. This doubt comes from the fact that λ is not a linear parameter, and its estimation is not part of the least squares fit.

Exercises

Please attempt the following exercises. Recall that to attempt these questions, you need to install the `faraway` package first.

Exercise 1

For the `prostate` data, fit a model with `lpsa` as the response and the other variables as predictors.

- (a) Suppose a new patient with the following values arrives:

$$\begin{aligned} \texttt{lcavol} &= 1.44692, \\ \texttt{lweight} &= 3.62301, \\ \texttt{age} &= 65, \\ \texttt{lbph} &= 0.30010, \\ \texttt{svi} &= 0.00000, \\ \texttt{lcp} &= -0.79851, \\ \texttt{gleason} &= 7.00000, \\ \texttt{pgg45} &= 15.00000. \end{aligned}$$

  Predict the `lpsa` for this patient along with an appropriate 95% CI.

- (b) Repeat the last question for a patient with the same values except that he is age 20. Explain why the CI is wider.

- (c) For the model of the previous question, remove all the predictors that are not significant at the 5% level. Now recompute the predictors. Which predictors would you prefer? Explain.

Exercise 2

Using the `teengamb` data, fit a model with `gamble` as the response and the other variables as predictors.

- (a) Predict the amount that a male with average (given these data) status, income, and verbal score would gamble along with an appropriate 95% CI.

- (b) Repeat the prediction for a male with maximal values (for this data) of status, income, and verbal score. Which CI is wider, and why is this result expected?

- (c) Fit a model with $\sqrt{\texttt{gamble}}$ as the response but with the same predictors. Now predict the response and give a 95% prediction interval for the individual in (a). Take care to give your answer in the original units of the response.
- (d) Repeat the prediction for the model in (c) for a female with `status`=20, `income`=1, `verbal`=10. Comment on the credibility of the result.

References

- Faraway, J. (2015). Linear Models with R Second Edition CHAPMAN & HALL/CRC Texts in Statistical Science.

Chapter 8

Appendix A: Linear Algebra

8.1 Terminology

In this lecture notes a **vector**[1] is always a column vector, this will be denoted as

$$\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ . \\ . \\ . \\ a_n \end{pmatrix}$$

the transpose of a column vector is a row vector and is denoted as $\mathbf{a}^\top = (a_1, a_2, \ldots, a_n)$. A matrix can be defined a rectangular array. A matrix $\mathbf{A} \in \mathbb{R}^{n \times k}$ can be written as

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1k} \\ a_{21} & a_{22} & \ldots & a_{2k} \\ a_{31} & a_{32} & \ldots & a_{3k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nk} \end{pmatrix}$$

the first index of the element a_{ij} refers to the ith row, and the second index to the jth column. The symbol $^\top$ denotes the transpose of a matrix $\mathbf{A} \in \mathbb{R}^{k \times n}$

[1] vector and matrices are denoted in bold

$$\mathbf{A}^\top = \begin{pmatrix} a_{11} & a_{21} & \dots & a_{n1} \\ a_{12} & a_{22} & \dots & a_{n2} \\ a_{13} & a_{23} & \dots & a_{n3} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1k} & a_{2k} & \dots & a_{nk} \end{pmatrix}$$

where the column of $\mathbf{A}$ are the rows of $\mathbf{A}^\top$, and vice versa. A matrix is said to be a **square matrix** if $n = k$. A square matrix is **symmetric** is $\mathbf{A} = \mathbf{A}^\top$, this implies that $a_{ij} = a_{ji}$. An example of square matrix in statistics is the covariance matrix. A square matrix $\mathbf{S}$ looks like as

$$\mathbf{S} = \begin{pmatrix} 1 & 0.5\ (a_{12}) & 0.7\ (a_{13}) \\ 0.5\ (a_{21}) & 1 & 0.3\ (a_{23}) \\ 0.7\ (a_{31}) & 0.3\ (a_{32}) & 1 \end{pmatrix}$$

where in brackets we denoted the ijth element of the matrix. It follows that its transpose is

$$\mathbf{S}^\top = \begin{pmatrix} 1 & 0.5\ (a_{21}) & 0.7\ (a_{31}) \\ 0.5\ (a_{12}) & 1 & 0.3\ (a_{32}) \\ 0.7\ (a_{13}) & 0.3\ (a_{23}) & 1 \end{pmatrix}$$

which is the exact same matrix but with indexes swapped.

A square matrix is called **diagonal** denoted with $\mathbf{D}$ if $a_{ij} = 0$ for all $i \neq j$

$$\mathbf{D} = \begin{pmatrix} a_{11} & 0 & \dots & 0 \\ 0 & a_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & a_{nk} \end{pmatrix}$$

the **identity** matrix $\mathbf{I}$ is a diagonal matrix with all diagonal elements equal to one.

$$\mathbf{I} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}$$

8.2 Properties

Let consider a number of vectors $\mathbf{a}_1$ to $\mathbf{a}_k$, we can take a linear combination of these vectors. With scalar weights $c_1, \dots, c_k$ this produces the vector $c_1\mathbf{a}_1 + c_2\mathbf{a}_2 + \cdots + c_k\mathbf{a}_k$, which we can shortly write as $\mathbf{Ac}$, where $\mathbf{A} \in \mathbb{R}^{n \times k}$ than can be written as $\mathbf{A} = (\mathbf{a}_1, \dots, \mathbf{a}_k)$ and $\mathbf{c} = (c_1, \dots, c_k)^\top$

A set of linear vectors is linearly dependent if any of the vectors can be written as a linear

combination of the others. That is, if there exist values for $c_1, \ldots, c_k$ not all zero, such that $c_1\mathbf{a}_1 + \cdots + c_k\mathbf{a}_k = 0$. Equivalently, a set of vectors is linearly independent if the only solution to

$$c_1\mathbf{a}_1 + \cdots + c_k\mathbf{a}_k = \mathbf{0}$$

is the trivial solution

$$c_1 = c_2 = \cdots = c_k = 0$$

That is, if the only solution to $\mathbf{Ac} = \mathbf{0}$ is $\mathbf{c} = \mathbf{0}$, where $\mathbf{0} = (0, \ldots, 0)^\top$ is the null vector.

If we consider all possible vectors that can be obtained as linear combinations of the vectors $\mathbf{a}_1, \ldots, \mathbf{a}_k$, these vectors form a vector space. If the vectors $\mathbf{a}_1, \ldots, \mathbf{a}_k$ are linearly dependent, we can reduce the number of vectors without changing this vector space. The minimal number of vectors needed to span a vector space is called the dimension of that space. This way, we can define the column space of a matrix as the space spanned by its columns, and the column rank of a matrix as the dimension of its column space. Clearly, the column rank can never exceed the number of columns. A matrix is of full column rank if the column rank equals the number of columns. The row rank of a matrix is the dimension of the space spanned by the rows of the matrix. In general, it holds that the row rank and the column rank of a matrix are equal, so we can unambiguously define the rank of a matrix. Note that this does not imply that a matrix that is of full column rank is automatically of full row rank (this is true only if the matrix is square). A useful result in regression analysis is that for any matrix $\mathbf{A}$

$$rank(\mathbf{A}) = rank(\mathbf{AA}^\top) = rank(\mathbf{A}^\top\mathbf{A}).$$

8.3 Inverse Matrix

A matrix $\mathbf{B}$, if it exists, is the **inverse** of a matrix $\mathbf{A}$ if $\mathbf{AB} = \mathbf{BA} = \mathbf{I}$. A necessary requirement fot his is that $\mathbf{A}$ is a square matrix and has full rank. In this case $\mathbf{A}$ is also called invertible or nonsingular. In this case we define $\mathbf{B} = \mathbf{A}^{-1}$, and

$$\mathbf{AA}^{-1} = \mathbf{I} \text{ and } \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}.$$

Note that the definition implies that $\mathbf{A} = \mathbf{B}^{-1}$. Thus we have $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$. If $\mathbf{A}^{-1}$ does not exist, we say that $\mathbf{A}$ is singular.

Suppose we are asked to solve $\mathbf{Ac} = \mathbf{d}$ where $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{c} \in \mathbb{R}^n$ and $\mathbf{d} \in \mathbb{R}^n$. This is a system of n linear equations with n unknowns. If $\mathbf{A}^{-1}$ exists, we can write

$$\mathbf{A}^{-1}\mathbf{Ac} = \mathbf{c} = \mathbf{A}^{-1}\mathbf{d}$$

to obtain the solution. If $\mathbf{A}$ is not invertible, the system of linear equations has linear depen-

dencies. Two possibilities can happen. Either more than one vector $\mathbf{c}$ satisfies $\mathbf{Ac} = \mathbf{d}$, so no unique solution exists, or the solutions are inconsistent, so there is no solution to the system. If $\mathbf{d} \equiv \mathbf{0}$, only the first possibility remains.

It is straightforward to derive that

$$(\mathbf{A}^{-1})^\top = (\mathbf{A}^\top)^{-1}$$

and

$$(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$$

8.4 Idempotent Matrices

A special class of matrices is that of symmetric and idempotent. A matrix $\mathbf{P}$ is symmetric if $\mathbf{P}^\top = \mathbf{P}$ and idempotent if $\mathbf{PP} = \mathbf{P}$. A symmetric idempotent matrix $\mathbf{P}$ has the interpretation of a projection matrix. This means that the projection vector $\mathbf{Px}$ is in the column space of $\mathbf{P}$, while the residual vector $\mathbf{x} - \mathbf{Px}$ is orthogonal to any vector in the column space of $\mathbf{P}$.

A projection matrix that projects upon the column space of a matrix $\mathbf{A}$ can be constructed as $\mathbf{P} = \mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top$. Clearly if we try to project twice on the same space we end up with no effect of the original projection, $\mathbf{PPx} = \mathbf{Px}$, *try to attempt this result, as exercise.* The residual from the projection matrix with $\mathbf{MP} = \mathbf{PM} = \mathbf{0}$ and $\mathbf{MM} = \mathbf{M} = \mathbf{M}^\top$. Thus the vectors $\mathbf{Mx}$ and $\mathbf{Px}$ are orthogonal. The only nonsingular projection matrix is the identity matrix. All other projection matrices are singular, each having rank equal to the dimension of the space upon which they project.

8.5 Eigenvalues and Eigenvectors

Let $\mathbf{A}$ be a symmetric $n \times n$ matrix. Consider the following problem of finding combinations of a vector $\mathbf{c}$ (other than the trivial solution) and a scalar that satisfy

$$\mathbf{Ac} = \lambda\mathbf{c}$$

At a first glance, this could seems a useless problem to pose for a statistician. However, what we are trying to do is to compress all the information that is incorporated in $\mathbf{A}$ into λ. This means that once the solutions of the system are obtained, if they exists, are the only ingredients needed to reconstruct $\mathbf{A}$. In general there are n solutions $\lambda_1, \ldots, \lambda_n$, called the eigenvalues of $\mathbf{A}$, corresponding to n vectors $c_1, \ldots, c_n$, called eigenvalues. If c_1 is a solution, the eigenvectors are defined up to a constant. The eigenvectors of a symmetric matrix are orthogonal, that is, $\mathbf{c}_i^\top\mathbf{c}_j = 0$

for all $i \neq j$. A singular matrix has at least one zero eigenvalue. In general, the rank of a symmetric matrix corresponds to the number of nonzero eigenvalues.

A symmetric matrix is called positive definite if all its eigenvalues are positive. It is called positive semi-definite if all its eigenvalues are non- negative, A positive definite matrix can be inverted. If $\mathbf{A}$ is positive definite, it holds for any vector $\mathbf{x}$ that

$$\mathbf{x}^\top \mathbf{A}\mathbf{x} > 0$$

The determinant of a symmetric matrix equals the product of its n eigenvalues. The determinant of a positive definite matrix is positive. A symmetric matrix is singular if the determinant is zero.

8.6 Differentiation

Let $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{c} \in \mathbb{R}^n$, then $\mathbf{c}^\top \mathbf{x}$ is a scalar. Let us consider $\mathbf{c}^\top \mathbf{x}$ as a function of $\mathbf{x}$. Then we can take the derivative respect to each of the elements in $\mathbf{x}$

$$\frac{\partial \mathbf{c}^\top \mathbf{x}}{\partial \mathbf{x}} = \mathbf{c}$$

this is a column vector of n derivatives. More generally for a vectorial function $\mathbf{A}\mathbf{x}$ where $\mathbf{A}$ is a matrix

$$\frac{\partial \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \mathbf{A}^\top$$

where the element in the column i, row j of this matrix is the derivative of the jth element in the function $\mathbf{A}\mathbf{x}$ with respect to x_i. Further

$$\frac{\partial \mathbf{x}^\top \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{A}\mathbf{x}$$

for a symmetric matrix.

If $\mathbf{A}$ is not symmetric, we have

$$\frac{\partial \mathbf{x}^\top \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = (\mathbf{A} + \mathbf{A}^\top)\mathbf{x}$$

8.7 Expectation and Covariance Operators

Let X_{ij}, $i = 1, 2, \ldots, m; j = 1, 2, \ldots, n$ be a set of random variables with expectation $\mathbb{E}(X_{ij})$. Whose in matrix form becomes $\mathbb{E}(\mathbf{X})$

$$\mathbb{E}(\mathbf{X}) = \begin{pmatrix} \mathbb{E}(X_{11}) & \mathbb{E}(X_{12}) & \dots & \mathbb{E}(X_{1n}) \\ \mathbb{E}(X_{21}) & \mathbb{E}(X_{22}) & \dots & \mathbb{E}(X_{2n}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}(X_{m1}) & \mathbb{E}(X_{m2}) & \dots & \mathbb{E}(X_{mn}) \end{pmatrix}$$

Theorem 4. *If* $\mathbf{A} \in \mathbb{R}^{l\times m}$, $\boldsymbol{B} \in \mathbb{R}^{n\times p}$ *and* $\mathbf{C} \in \mathbb{R}^{l\times p}$ *matrices, respectively of constants, then*

$$\mathbb{E}[\mathbf{A}\mathbf{X}\boldsymbol{B} + \mathbf{C}] = \mathbf{A}\mathbb{E}[\mathbf{X}]\boldsymbol{B} + \mathbf{C}$$

(*Proof* in Seber (2003))this important results holds also if $\mathbf{X} \in \mathbb{R}^m$ (m dimensional vector) then $\mathbb{E}[\mathbf{A}\mathbf{X}] = \mathbf{A}\mathbb{E}[\mathbf{X}]$. Let $\mathbf{A}$ and $\boldsymbol{B}$ be $m \times n$ matrices of constants, and $\mathbf{X}$ and $\mathbf{Y}$ be $n \times 1$ vectors of random variables, then

$$\mathbb{E}[\mathbf{A}\mathbf{X} + \boldsymbol{B}\mathbf{Y}] = \mathbf{A}\mathbb{E}[\mathbf{X}] + \boldsymbol{B}\mathbb{E}[\mathbf{Y}].$$

In a similar manner we can generalize the notions of covariance and variance for vectors. If $\mathbf{X} \in \mathbb{R}^m$ and $\mathbf{Y} \in \mathbb{R}^n$ vectors of random variables, then we define the generalized covariance operator $Cov(\cdot)$ as

Theorem 5. *If* $\mathbb{E}(\mathbf{X}) = \mu_X$ *and* $\mathbb{E}(\mathbf{Y}) = \mu_Y$ *then*

$$Cov[\mathbf{X}, \mathbf{Y}] = \mathbb{E}[(\mathbf{X} - \mu_X)(\mathbf{Y} - \mu_Y)]$$

When $\mathbf{X} = \mathbf{Y}$, $Cov[\mathbf{X}\mathbf{X}]$ written as $Var[\mathbf{X}]$ is called the variance (variance-covariance or dispersion) matrix of $\mathbf{X}$

$$Var(\mathbf{X}) = \begin{pmatrix} Var(X_1) & Cov(X_1, X_2) & \dots & Cov(X_1, X_n) \\ Cov(X_2, X_1) & Var(X_2) & \dots & Cov(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(X_n, X_1) & Cov(X_n, X_2) & \dots & Var(X_n) \end{pmatrix}$$

Since $Cov[X_iX_j] = Cov[X_jX_i]$, the matrix above is symmetric.

From Theorem **??**, with $\mathbf{Y} = \mathbf{X}$ we have

$$Var[\mathbf{X}] = \mathbb{E}[(\mathbf{X} - \mu_X)(\mathbf{X} - \mu_X)^\top]$$

which, after an expansion, leads to

$$Var[\mathbf{X}] = \mathbb{E}[\mathbf{X}\mathbf{X}^\top] - \mu_X\mu_X^\top$$

These last two equations are natural generalization of univariate results

Theorem 6. *If $\mathbf{X} \in \mathbb{R}^m$ and $\mathbf{Y} \in \mathbb{R}^n$ be vectors of random variables, and $\mathbf{A}$ and $\mathbf{B}$ are $l \times m$ and $p \times n$ matrices of constants, then*

$$Cov[\mathbf{AX}, \boldsymbol{B}\mathbf{Y}] = \mathbf{A}Cov[\mathbf{X}, \mathbf{Y}]\mathbf{B}^\top$$

(*proof in Seber (2003)*), from the Theorem just stated we have the special cases

$$Cov(\mathbf{AX}, \mathbf{Y}) = \mathbf{A}Cov(\mathbf{X}, \mathbf{Y})$$

$$Cov(\mathbf{X}, \boldsymbol{B}\mathbf{Y}) = Cov(\mathbf{X}, \mathbf{Y})\boldsymbol{B}^\top$$

Of particular importance is the following result, obtained by setting $\boldsymbol{B}^\top = \mathbf{A}$ and $\mathbf{Y} = \mathbf{X}$

$$Var[\mathbf{AX}] = Cov[\mathbf{AX}, \mathbf{AX}] = \mathbf{A}Cov[\mathbf{XX}]\mathbf{A}^\top = \mathbf{A}Var[\mathbf{X}]\mathbf{A}^\top$$

Theorem 7. *If $\mathbf{X}$ is a vector of random variables such that no elements of $\mathbf{X}$ is a linear combination of the remaining elements [i.e. there do not exist $\mathbf{a}(\neq \mathbf{0})$ and b such that $\mathbf{a}^\top\mathbf{X} = b$ for all values of $\mathbf{X} = \mathbf{x}$], then $Var[\mathbf{X}]$ is a positive definite matrix.*

(*Proof in Seber (2003)*).

Quadratic forms play a major role in multivariate linear regression. In particular, we will frequently need to find the expected value of a quadratic form using the following theorem.

Theorem 8. *Let $\mathbf{X} \in \mathbb{R}^n$ vector of random variables, and let $\mathbf{A} \in \mathbb{R}^{n\times n}$ symmetric matrix. If $\mathbb{E}[\mathbf{X}] = \boldsymbol{\mu}$ and $Var[\mathbf{X}] = \boldsymbol{\Sigma}$ then*

$$\mathbb{E}[\mathbf{X}^\top\mathbf{AX}] = tr(\mathbf{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^\top\mathbf{A}\boldsymbol{\mu}$$

Proof.

$$\begin{aligned}
\mathbb{E}[\mathbf{X}^\top\mathbf{AX}] &= tr(\mathbb{E}[\mathbf{X}^\top\mathbf{AX}]) \\
&= \mathbb{E}[tr(\mathbf{X}^\top\mathbf{AX})] \\
&= \mathbb{E}[tr(\mathbf{AXX}^\top)] \\
&= tr(\mathbb{E}[\mathbf{AXX}^\top]) \\
&= tr(\mathbf{A}\mathbb{E}[\mathbf{XX}^\top]) \\
&= tr(\mathbf{A}(Var[\mathbf{X}] + \boldsymbol{\mu}\boldsymbol{\mu}^\top)) \\
&= tr(\mathbf{A}\boldsymbol{\Sigma}) + tr(\mathbf{A}\boldsymbol{\mu}\boldsymbol{\mu}^\top) \\
&= tr(\mathbf{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^\top\mathbf{A}\boldsymbol{\mu}
\end{aligned}$$

We can deduce two special cases. Fist, by setting $\mathbf{Y} = \mathbf{X} - \boldsymbol{B}$ and noting that $Var[\mathbf{Y}] = Var[\mathbf{X}]$ we have

$$\mathbb{E}[(\mathbf{X} - \boldsymbol{\mu})^\top \mathbf{A}(\mathbf{X} - \boldsymbol{\mu})] = tr(\mathbf{A}\boldsymbol{\Sigma}) + (\boldsymbol{\mu} - \boldsymbol{B})^\top \mathbf{A}(\boldsymbol{\mu} - \boldsymbol{B})$$

Second, if $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$ then $tr(\mathbf{A}\boldsymbol{\Sigma}) = \sigma^2 tr(\mathbf{A})$. Thus

$$\mathbb{E}[\mathbf{X}^\top \mathbf{A}\mathbf{X}] = \sigma^2(\text{sum of coefficients of } X_i^2) + (\mathbf{X}^\top \mathbf{A}\mathbf{X})_{\mathbf{x}=\boldsymbol{\mu}}$$

Chapter 9

Appendix B: Multivariate Normal Distribution

The multivariate normal distribution plays a fundamental role in linear regression. While real data are never exactly multivariate normal, the normal density is often a useful approximation to the "true" population distribution.

The main advantage of multivariate normal distribution lies in the fact that it is mathematically tractable and "nice" results can be derived and studied from it. This is frequently not the case for other distributions expecially in the multivariate domain. Of course, mathematical attractiveness per se is of little use to the practitioner. It turns out, however, that normal distributions are useful in practice for two reasons:

- First, the normal distribution serves as a bona fide population model in some instances;
- Second, the sampling distributions of many multivariate statistics are approximately normal, regardless of the form of the parent population, because of a central limit effect.

The multivariate normal density is a generalization of the univariate normal density when $p \geq 2$. Recall that the univariate normal distribution, with mean μ and variance σ^2, can be written as

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-[(x-\mu)/\sigma]^2/2} \quad -\infty < x < \infty$$

It is convenient to denote the normal density function with mean μ and variance σ^2 by $N(\mu, \sigma^2)$. The term

$$\left(\frac{x-\mu}{\sigma}\right)^2 = (x-\mu)\left(\sigma^2\right)^{-1}(x-\mu)$$

in the exponent of the univariate normal density function measures the square of the distance from x to μ in standard deviation units. This can be generalized for a $p \times 1$ vector $\mathbf{x}$ of observations

on several variables as

$$(\mathbf{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$$

The $p \times 1$ vector $\boldsymbol{\mu}$ represents the expected value of the random vector $\mathbf{X}$, and the $p \times p$ matrix $\boldsymbol{\Sigma}$ is the variance-covariance matrix of $\mathbf{X}$. We shall assume that the symmetric matrix $\boldsymbol{\Sigma}$ is positive definite, so the expression in is the square of the generalized distance from $\mathbf{x}$ to $\boldsymbol{\mu}$.

The multivariate normal density is obtained by replacing the univariate distance by the multivariate generalized distance of in the density function. When this replacement is made, the univariate normalizing constant $(2\pi)^{-1/2}\left(\sigma^2\right)^{-1/2}$ must be changed to a more general constant that makes the volume under the surface of the multivariate density function unity for any p. This is necessary because, in the multivariate case, probabilities are represented by volumes under the surface over regions defined by intervals of the x_i values. It can be shown that this constant is $(2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2}$, and consequently, a p-dimensional normal density for the random vector $\mathbf{X}^{\top} = [X_1, X_2, \ldots, X_p]$ has the form

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-(\mathbf{x}-\mu)^{\top}\Sigma^{-1}(\mathbf{x}-\mu)/2}$$

where $-\infty < x_i < \infty, i = 1, 2, \ldots, p$. We shall denote this p-dimensional normal density by $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, which is analogous to the normal density in the univariate case.

The key properties of a Multivariate normal random variable $\mathbf{X}$ are:

1. Linear combinations of the components of $\mathbf{X}$ are normally distributed.
2. All subsets of the components of $\mathbf{X}$ have a (multivariate) normal distribution.
3. Zero covariance implies that the corresponding components are independently distributed.
4. The conditional distributions of the components are (multivariate) normal.