Theory of Classical Statistics II
Regression and Multivariate Analysis
with demonstration in Julia

Samuel Lam
Imperial College // Massachusetts Institute of Technology

Spring 2021

These notes are taken based on the notes from the following lectures.

- (Imperial) M2AA3 Introduction to Numerical Analysis
- (Imperial) M2S1 Probability and Statistics II
- (Imperial) M2S2 Statistical Modelling I
- (Imperial) M3S2 Statistical Modelling II
- (Imperial) M4S18A2 Multivariate Analysis
- (MIT) 18.338 Random Matrix Theory
- (MIT, OCW) 18.650 Statistics for Applications
- (MIT, OCW) 18.655 Mathematical Statistics
- (Cambridge) Part IB Statistics

The notes are not endorsed by the lecturer, and are significantly modified to ensure coherence among other related courses. The author welcomes any suggestions - please email your suggestions to `chun.lam18@imperial.ac.uk`.

This series of notes are dedicated to (in alphabetical order)

- the *Almighty God*
- Author's high school teachers - Mr. Vincent Kong and Mr. Kenny Koon
- Author's friends - Cyrus Chan, Jeromy Leong, Ciaran Leung, Derek Leung, Marcus Law, Carlos Wong.

First printing, January 2021

Contents

Regression

CHAPTER 1

Introduction

1.1 Regression

A fundamental building block of modern statistical theory is the study of relationship between a (*response*) random variable and one or more *co-variates* (independent random variables). Specifically, let $(\vec{X}, Y) = (X^{(1)}, ..., X^{(d)}, Y)$ be a $d+1$-dimensional vector of random variables (or *random vector*) defined on a certain sample space [1] $(\Omega, \mathcal{F}, \mathbb{P})$. What can we say about the conditional expectation $\mathbb{E}(Y \mid \vec{X})$, if we are given n samples of the random variable $(\vec{X}, Y)$, say $\{(\vec{x}_i, y_i)\}_{i=1}^n := \left\{(x_i^{(1)}, ..., x_i^{(d)}, y_i)\right\}_{i=1}^n$?

There are various forms of $\mathbb{E}(Y \mid \vec{X})$. For instance, the *linear model* takes the form

$$\mathbb{E}(Y \mid \vec{X}) = \beta_1 f_1(\vec{X}) + ... + \beta_p f_p(\vec{X}), \quad g : \mathbb{R} \to \mathbb{R}, \ f_j : \mathbb{R}^p \to \mathbb{R} \tag{1.1}$$

with β_i unknown (but fixed) parameters to be estimated, corresponding to the functions f_i. For the model to be well-defined, we want the $f_1, ..., f_p$ to be linearly independent [2] This is known as the *Full Rank* (FR) assumption.

A more general version of linear model, the *Generalised Linear Model (GLM)*, takes the form

$$g(\mathbb{E}(Y \mid \vec{X})) = \beta_1 f_1(\vec{X}) + ... + \beta_p f_p(\vec{X}), \quad g : \mathbb{R} \to \mathbb{R}, \ f_j : \mathbb{R}^p \to \mathbb{R} \tag{1.2}$$

where g is known as a *link* function and again f_i are linearly independent. If we assume these models *a priori* then we can understand $\mathbb{E}(Y \mid X)$ by estimating the parameters $\beta_1, ..., \beta_p$, i.e. reducing to a parameter estimating problem. To avoid confusion, let us keep track of the variables we have used – there are d covariates, p unknown parameters and we obtain samples of size n from it. In practice we want $n \geq p$. [3]

[1] we don't need to know what exactly it is

[2] which means if $\lambda_1, ..., \lambda_p \in \mathbb{R}$ satisfies $\lambda_1 f_1 + ... + \lambda_p f_p \equiv 0$ then $\lambda_1 = ... = \lambda_p = 0$.

[3] There will be identifiability issue if $n < p$, but we usually have large data set so this is rare.

Example 1.1.1: Bivariate Normal Model

Consider the bivariate normal model [a] which assumes

$$(X, Y) \sim \mathbf{N}_2 \left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix} \right) \tag{1.3}$$

This can be written in the form of (1.1) by noting that

$$\mathbb{E}(Y \mid X) = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(X - \mu_X) = \left(\mu_Y - \rho\frac{\sigma_Y}{\sigma_X}\mu_X\right) + \rho\frac{\sigma_Y}{\sigma_X}X \tag{1.4}$$

[a] we will define in next section in case you haven't seen this

Exercise 1.1.2

Show equation (1.4)

We are not done with defining a linear model / GLM yet. A standard procedure of analysing estimator (statistic) is to look at the distribution of estimator when 'evaluated' by the i.i.d. copies of $(\vec{X}, Y)$, say $\left\{(\vec{X}_i, Y_i)\right\}_{i=1}^n$. To use the brilliant tools in statistical theory and linear algebra, we need to simplify the definitions little bit: consider the very long random variable $(\vec{X}_1, \vec{X}_2, ..., \vec{X}_n, \vec{Y})$ with $\vec{Y} = (Y_1, ..., Y_n)$. We may then rewrite (1.1) as [4]

$$\mathbb{E}(\vec{Y} \mid \vec{X}_1, \ ... \vec{X}_n) = \underbrace{\begin{pmatrix} f_1(\vec{X}_1) & f_2(\vec{X}_1) & ... & f_p(\vec{X}_1) \\ f_1(\vec{X}_2) & f_2(\vec{X}_2) & ... & f_p(\vec{X}_2) \\ & & ... & \\ f_1(\vec{X}_n) & f_2(\vec{X}_n) & ... & f_p(\vec{X}_n) \end{pmatrix}}_{:=X} \underbrace{\begin{pmatrix} \beta_1 \\ \beta_2 \\ ... \\ \beta_p \end{pmatrix}}_{:=\vec{\beta}} = X\vec{\beta} \tag{1.5}$$

which we abuse notation and write X as (known as *design matrix*) as above.

Remark.

- It is immediately clear why the assumption on f_j's is called the Full Rank assumption – if this assumption holds then the matrix X has full rank.
- A heuristic way to check if the assumption is *valid* for a particular dataset $\left\{(x_i^{(1)}, ..., x_i^{(d)}\right\}_{i=1}^n$ is to plug in the x values into X and see if it is full rank.
- There are issues even when the columns of X is highly correlated. To mitigate with this we attempt the drop some of the f_j's by techniques like *Ridge Regression* – to be discussed later in this book.

Notice that the random variables $\epsilon_i := Y_i - \mathbb{E}(Y_i \mid \vec{X}_i)$ should be i.i.d., independent from X_i and has mean zero. For our convenience we also want ϵ_i to have at least 2nd moment, i.e. $\text{Var}(Y_i) = \sigma^2 < \infty$. These assumptions are often called Second-Order Assumption (SOA). We may then properly define the linear model.

[4] Expectation is defined componentwise.

Definition 1.1.3: (Normal) Linear Model

A linear model assumes the joint distribution of $(\vec{X}_1, \vec{X}_2, ..., \vec{X}_n, \vec{Y})$ with $\vec{Y} = (Y_1, ..., Y_n)$ satisfies

$$\vec{Y} \mid \vec{X}_1, ..., \vec{X}_n \sim X\vec{\beta} + \vec{\epsilon} \tag{1.6}$$

where X and $\vec{\beta}$ are specified in (1.5), f_j's satisfies (FM), $\text{Cov}(\vec{\epsilon}) = \sigma^2 I^{(n)}$ [a] with $I^{(n)}$ $n \times n$ identity matrix. In particular, if $\vec{\epsilon} \sim \mathbf{N}_n(\vec{0}, \sigma^2 I^{(n)})$ (i.e. is multivariate normally distributed) [b], then this model is known as a Normal linear model.

[a] See later section for definition, in case you haven't seen this
[b] this is the Normal Theory Assumption (NTA)

Remark. We may heuristically interpret the conditional distribution by treating $\vec{X}_1, ..., \vec{X}_n$ as constants (hence also X). Sometimes we might even drop the condition sign.

Exercise 1.1.4: Linear Models?

Determine if the following are linear model. (assume all ϵ_i's are iid $\mathbf{N}(0, 1)$.)

1. $Y_i = \mu + \epsilon_i$
2. $Y_i = \begin{cases} \mu_1 & i = 1, ..., m \\ \mu_2 & i = m+1, ..., n \end{cases}$
3. $Y_i = a + b \ln X_i + \epsilon_i$
4. $\ln Y_i = a + b \ln X_i + \epsilon_i$
5. $Y_i = aX_i^2 + bX_i + c + \epsilon_i$
6. $Y_i = a(X_i - h)^2 + k + \epsilon_i$

Hint. Yes for (1), (2), (3) and (5). No for (4) and (6). (1) is arguable the simplest linear model, and has much significance! (2) represents the model in ANOVA (analysis of variance). It is interesting to see that (5) and (6) both represents a quadratic model, but (6) is much more harder than (5) to work with.

The SOA is included in our definition of linear model to avoid correlation in between samples (particularly the Y_i's.) This often arises when we consider X_i as times. In such case the error ϵ_i might have much more complicated structure (e.g. moving averages), requiring more sophisticated tool to handle.

Remark. In the case when $\text{Cov}(\vec{\epsilon}) = \sigma^2 V$, where V is a symmetric positive-definite (spd) matrix, then V has a spectral decomposition $V = Q\Lambda Q^T$ with Q orthogonal [a] and Λ diagonal. Let $T = Q\Lambda^{-1/2}Q^T$, then

$$T^T V T = Q\Lambda^{-1/2}Q^T Q\Lambda Q^T Q\Lambda^{-1/2}Q^T = I^{(n)}$$
$$TT^T = Q\Lambda^{-1/2}Q^T Q\Lambda^{-1/2}Q^T = V^{-1}$$

then by introducing $\vec{Y}' = T^T\vec{Y}$ then SOA is satisfied. This trick is called weighted least squares (WLS) which we won't go into details.

[a] Q is orthogonal if $Q^TQ = QQ^T = I^{(n)}$

1.2 Data Example I - Heights of Father and Son

File: `PearHeight.csv`

Our first dataset would be Karl Pearson's experiments on correlations among heights of various members in a family in 1903. For each of 1078 families, the heights of various family members. How would the heights of sons depend on the heights on fathers?

1.2.1 Reminder of Important Julia Codes

Before that we need to import the data. The datasets we use in this book always have a Comma Seperated Values (`.csv`) format, and the `CSV` library will be able to handle these files when using with the `DataFrames` library.

```
using CSV, DataFrames
```

Now we can import the data using the following command:

```
data = CSV.read("(path of your file)", DataFrame)
```

Of course, if you save the dataset in the same directory with your julia code, then you can use the relative path, i.e.

```
data = CSV.read("./PearHeight.csv", DataFrame)
```

The first few line of your output should be as followed:

```
1078x2 DataFrame
| Row  | Father  | Son     |
|      | Float64 | Float64 |
|------|---------|---------|
| 1    | 65.0    | 59.8    |
| 2    | 63.3    | 63.2    |
| 3    | 65.0    | 63.3    |
...
```

Notice the variable `Data` has type `DataFrame`. This datatype is the one we usually use to handle data. Notice that this particular `DataFrame` has two fields (as suggested by the first row) - `Father` and `Son`. We can call the data in one field (say `Father`) using the following command:

```
data.Father
```

which returns the following

```
1078-element Array{Float64,1}:
 65.0
 63.3
 65.0
 ...
```

Notice `data.Father` has type `Array{Float64,1}`, which is easier to work with. You may also retrieve (and modify) one or more specific data entries (say entries 74, 262 and 751) using a similar way as how you normally retrieve entries of an `Array`: (we will later demonstrate the use of dot notations)

```
data[[74,262,751],:]
```

```
3x2 DataFrame
| Row | Father  | Son     |
|     | Float64 | Float64 |
|-----|---------|---------|
| 1   | 62.9    | 68.5    |
| 2   | 68.7    | 69.0    |
| 3   | 71.6    | 68.0    |
```

Finally, it might be easier for us to work with `Array` instead of `DataFrame`. We may convert an `DataFrame` to a `Array` by using the following command

```
convert(Matrix,data)
```

```
1078x2 Array{Float64,2}:
 65.0  59.8
 63.3  63.2
 65.0  63.3
...
```

We will see later that an `Array` to a `DataFrame` using the following command

```
DataFrame((the array to be converted))
```

Finally, you can append a column using the insert commands, e.g.

```
insertcols!(data, 3, :new_col => 1:1078 )
```

```
1078x3 DataFrame
| Row  | Father  | Son     | new_col |
|      | Float64 | Float64 | Int64   |
|------|---------|---------|---------|
| 1    | 65.0    | 59.8    | 1       |
| 2    | 63.3    | 63.2    | 2       |
| 3    | 65.0    | 63.3    | 3       |
...
```

and remove a column using either indexing, or the `select!` command, e.g.:

```
select!(data, Not(:new_col))
```

1.2.2 Part I - First Exploration

We may now plot a scatter plot against the data we have. Be sure to have the libraries like `StatsPlots` imported:

```
using StatsPlots
plt01a = scatter(data.Father,data.Son,...)        # Left Plot
plt01b = marginalhist(data.Father,data.Son,...)   # Right Plot
plt01 = plot(plt01a,plt01b,layout=(1,2),...)      # Merge
```

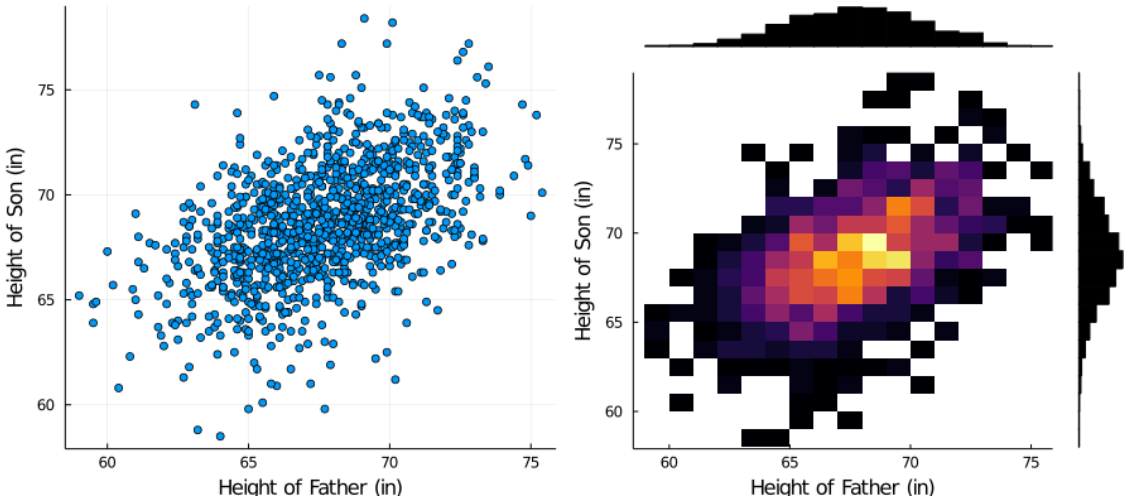

Figure 1.1: Scatter Plot of Data

Remark. You may add other attributes to the scatter plots, e.g. `xlabel`, `title`. The figure shown above use the following commands:

```
using StatsPlots   # or Plots
scatter(data.Father,data.Son,xlabel="Height of Father (in)",
 ylabel="Height of Son (in)",legend=false)
```

Only the essential parts of code are included.

The ellipse-like shape of cluster indicates that a Bivariate Normal Model (which is of a Normal Linear Model) might be a suitable model for the result of this experiment. Specifically, if $\{(X_i, Y_i)\}_{i=1}^{1078}$ are results of this experiment, then we know that each (X_i, Y_i)'s are i.i.d. and

$$\begin{pmatrix} X_i \\ Y_i \end{pmatrix} \sim \mathbf{N}_2 \left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix} \right)$$

As a notation we write $\{(x_i, y_i\}_{i=1}^{1078}$ as a (deterministic) *sample / realisation* of the distribution $\{(X_i, Y_i)\}_{i=1}^{1078}$. In our case we have $(x_1, y_1) = (65.0, 59.8)$, $(x_2, y_2) = (68.7, 69.0)$ etc. To assess the suitability of this model we may histogram a random linear combination of fathers' heights and sons' height.

```
using StatsPlots
a,b = rand(2)   # Generating Random Numbers
histogram(a.*data.Father + b.*data.Son)
```

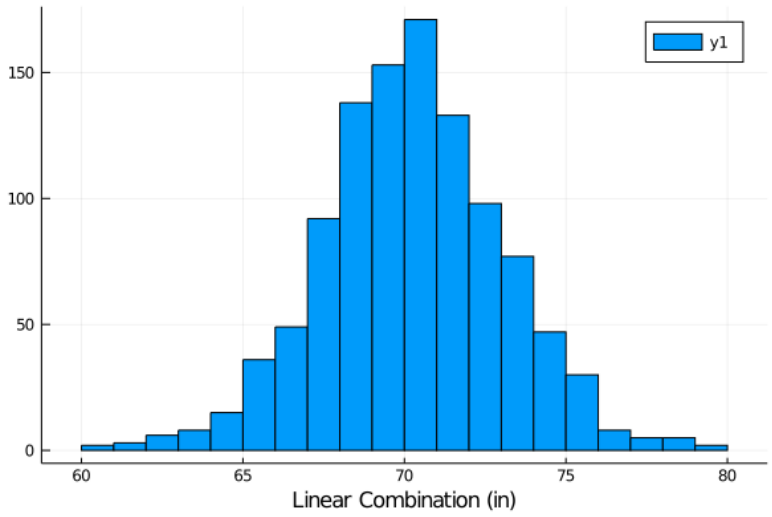

Figure 1.2: Histogram of Random Linear Combination of Heights

It has a bell shape, so it suggests that our bivariate normal model is suitable. There are several ways to test normality of a series of data $(z_i)_{i=1}^n$ (sorted in ascending order). If they are a realisation of i.i.d. normal distribution $Z_i \sim \mathbf{N}(\mu, \sigma^2)$, then by

law of large number the empirical distribution function (ECDF)

$$F_n(z) = \sum_{i=1}^{n} \mathbb{I}_{[z_i,\infty)}(z) \tag{1.7}$$

should "converge uniformly" [5] to the CDF of normal distribution $F(z) := F_{\mu,\sigma^2}(z) := \Phi((z-\mu)/\sigma)$ (with Φ CDF of standard normal distribution). In particular, $F_n(z_i) = i/n$ should be approximately equal to $F(z_i)$. Rearranging gives

$$z_i \approx F^{-1}(F_n(z_i)) = F^{-1}(i/n) \tag{1.8}$$

If we plot $F^{-1}(i/n)$ ("theoretical quantiles") against z_i ("actual quantiles"), we should get a straight line with slope 1 passing through the origin. In practice we replace μ with sample mean and σ^2 with sample variance. Here is an implementation of the plot.

```
using StatsPlots, Distributions
function my_qqnorm(samp)
    samp = sort(samp)                              # Sort Sample
    mu = mean(samp); sigma2 = std(samp)            # Mean and Variance
    invcdf(x) = quantile(Normal(mu,sigma2), x) # Quantiles
    # QQ Plot
    pl = scatter(invcdf.([i/length(samp) for i=1:length(samp)]),
    samp, label="sample", legend=:bottomright)

    # with the line y = x
    plot!(x -> x, invcdf(1/length(samp)),
    invcdf(1-1/length(samp)), label="y=x")

    # Further attributes
    xlabel!("Theoretical Quantiles")
    ylabel!("Actual Quantiles")
    return pl
end
```

We apply this code to our combination of heights:

```
a,b = rand(2)
z = a.*data.Father + b.*data.Son
my_qqnorm(z)
```

and obtain our Q-Q Plot. The scatter points lie on the line actual=theoretical, so we can say that bivariate normal model might be a good model. In fact, the `qqnorm` function in `StatsPlots` will produce the Q-Q Plot we need.

[5] See Glivenko Cantelli Theorem for more precise statement

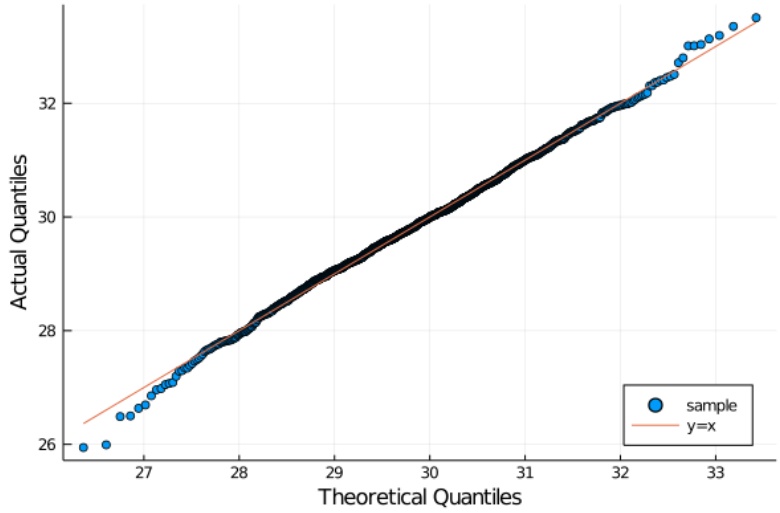

Figure 1.3: Q-Q Plot for Random Linear Combination of Heights

Remark. We can quantitatively assessing the normality of a series of data by *Anderson Darling Test*. Notice that the ordinary one-sided Kolmogorov-Smirnov Test is not valid since the parameters of our model are unknown. Here is a simple implementation:

```
using HypothesisTests
OneSampleADTest(z,fit(Normal,z))
```

```
One sample Anderson-Darling test
--------------------------------
Population details:
    parameter of interest:   not implemented yet
    value under h_0:         NaN
    point estimate:          NaN

Test summary:
    outcome with 95% confidence: fail to reject h_0
    one-sided p-value:           0.5804

Details:
    number of observations:   1078
    sample mean:              16.431648790072547
    sample SD:                0.6485448331246852
    A^2 statistic:            0.6745702853054821
```

Estimating μ_X and μ_Y in a Bivariate Normal Model is easy (consider sample mean), but estimating the covariance matrix is non-trivial (see later chapter). Instead we consider a simpler Normal Linear Model:

$$Y_i \,|\, X_i \sim \beta_1 + \beta_2 X_i + \epsilon_i, \quad \epsilon_i \sim \mathbf{N}(0, \sigma^2)$$

which can be rewritten in the form of (2.25) as followed (assuming $\vec{X} = (X_1, ..., X_{1078})$ and $\vec{Y} = (Y_1, ..., Y_{1078})$).

$$\vec{Y} \,|\, \vec{X} \sim \underbrace{\begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_{1078} \end{pmatrix}}_{:=X} \underbrace{\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}}_{:=\vec{\beta}} + \vec{\epsilon}, \qquad \vec{\epsilon} \sim \mathbf{N}_{1078}(0, \sigma^2 I^{(1078)}) \tag{1.9}$$

1.3 Review in Probability

I hope you have seen the following results in courses in elementary statistics – if not, the hints would be a guide for you to prove the results.

1.3.1 Random Vectors

Let us be reminded of some notions of random vectors [6], that are random variables $\vec{X} := (X_1, ..., X_n) : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ with n not necessary 1 – or a vector of n random variables $X_1, ..., X_n$ defined on the same *sample space* $(\Omega, \mathcal{F}, \mathbb{P})$. Note that any random vector comes with a (cumulative) *distribution*

$$F_{\vec{X}}(x_1, ..., x_n) = \mathbb{P}(X_1 \leq x_1, \, ..., \, X_n \leq x_n) \tag{1.10}$$

and (if exists by absolute continuity) a density

$$f_{\vec{X}}(x_1, ..., x_n) = \frac{\partial^n}{\partial x_1 \, ... \, \partial x_n} F_{\vec{X}}(\vec{x}) \tag{1.11}$$

Definition 1.3.1: Expectation

The expectation of a random vector is defined componentwise, so if $\vec{X} = (X_1, ..., X_n)$ is a random vector then $\mathbb{E}(\vec{X}) = (\mathbb{E}(X_1), ..., \mathbb{E}(X_n))$.

Exercise 1.3.2: Linearity of Expectation

Check that

- If $\vec{U}, \vec{V} \in \mathbb{R}^n$ are vectors of random variables, and $a \in \mathbb{R}$ is deterministic (not random), then $\mathbb{E}(\vec{V}^T) = (\mathbb{E}(\vec{V}))^T$, and $\mathbb{E}(a\vec{U} + \vec{V}) = a\mathbb{E}(\vec{U}) + \mathbb{E}(\vec{V})$
- Given $\vec{V} \in \mathbb{R}^n$ is vector of random variables, and $A \in \mathbb{R}^{r \times n}$, $B \in \mathbb{R}^{n \times c}$ are deterministic (i.e. not matrices of random variables), then $\mathbb{E}(A\vec{V}) = A\mathbb{E}(\vec{V})$ and $\mathbb{E}(\vec{V}^T B) = \mathbb{E}(\vec{V})^T B$.

Hint. By comparing components.

[6] we will be doing random matrices in second part

We also define covariance between two random vectors (which is a matrix).

Definition 1.3.3: Covariance Matrix

If $\vec{U} \in \mathbb{R}^n$ and $\vec{V} \in \mathbb{R}^m$, then the covariance matrix $\text{Cov}(\vec{U}, \vec{V})$ is defined componentwise as

$$(\text{Cov}(\vec{U}, \vec{V}))_{ij} = \text{Cov}((\vec{U})_i, (\vec{V})_j) \tag{1.12}$$

In addition, define $\text{Cov}(\vec{U}) = \text{Cov}(\vec{U}, \vec{U})$.

Exercise 1.3.4: Other Definitions of Covariance

$$\text{Cov}(\vec{U}, \vec{V}) = \mathbb{E}((\vec{U} - \mathbb{E}(\vec{U}))(\vec{V} - \mathbb{E}(\vec{V}))^T) = \mathbb{E}(\vec{U}\vec{V}^T) - \mathbb{E}(\vec{U})\mathbb{E}(\vec{V})^T \tag{1.13}$$

We can see that both Y_1 and Y_2 are normally ($\mathbf{N}(0, 1)$) distributed.

Hint. For first equality prove by comparing entries. Second equality can be proven by direct expansion of first equality.

We may therefore check that

Exercise 1.3.5

- If $\vec{U}, \vec{V} \in \mathbb{R}^n, \vec{W} \in \mathbb{R}^m$ are vectors of random variables, and $a \in \mathbb{R}$, then $\text{Cov}(a\vec{U} + \vec{V}, \vec{W}) = a\text{Cov}(\vec{U}, \vec{W}) + \text{Cov}(\vec{V}, \vec{W})$
- If $\vec{U}, \in \mathbb{R}^n, \vec{V} \in \mathbb{R}^m$ are vectors of random variables, and $A \in \mathbb{R}^{a\times n}, B \in \mathbb{R}^{b\times m}$, then $\text{Cov}(A\vec{U}, B\vec{V}) = A\text{Cov}(\vec{U}, \vec{V})B^T$.
- Let $\vec{V} \in \mathbb{R}^n$ be vector of random variables, then $\text{Cov}(\vec{V})$ is symmetric and non-negatively definite.

Hint. First two inequalities can be established by previous exercise. To show $\text{Cov}(\vec{V})$ is symmetric notice $\text{Cov}(\vec{U}, \vec{V}) = \text{Cov}(\vec{V}, \vec{U})^T$. For non-negative definite, think about $\vec{c}^T\text{Cov}(\vec{V})\vec{c}$ for arbitary $\vec{c} \in \mathbb{R}^n$.

1.3.2 Multivariate Normal Distribution

Perhaps the most important multivariate distribution we have is the multivariate normal distribution (MVN). The definition is as followed:

Definition 1.3.6: Multivariate Normal (Gaussian) Distribution

A random vector is $\vec{X} = (X_1, ..., X_n)$ is multivariate normally (MVN) distributed iff for all $\vec{\alpha} \in \mathbb{R}^n$ we have $\vec{\alpha}^T\vec{X}$ normally distributed.

A canonical example would be the *vector of standard Gaussian* $\vec{Z} = (Z_1, ..., Z_n)$ with Z_i iid $\mathbf{N}(0, 1)$. Then for all $\vec{\alpha} = (\alpha_1, ..., \alpha_n)$ we have $\vec{\alpha}^T\vec{X} = \mathbf{N}(0, \sum_i \alpha_i^2)$. Therefore $\vec{Z}$ is multivariate normally distributed with mean $\vec{0}$ and covariance $I^{(n)}$.

Remark. It is tempting to think that if all entries of a random vector $\vec{Y}$ are normally distributed then $\vec{Y}$ is multivariate normally distributed. However this is not true in general. Let $Y_1 \sim \mathbf{N}(0,1)$ and Z follows a symmetric Bernoulli distribution with $\mathbb{P}(Z=1) = \mathbb{P}(Z=-1) = 1/2$. Let $Y_2 = Y_1 Z$. Then

$$\begin{aligned} F_{Y_2}(y) = \mathbb{P}(Y_2 \leq y) &= \mathbb{P}(Y_2 \leq y \,|\, Z=1)\mathbb{P}(Z=1) + \mathbb{P}(Y_2 \leq y \,|\, Z=-1)\mathbb{P}(Z=-1) \\ &= \frac{1}{2}(\mathbb{P}(Y_1 \leq y) + \mathbb{P}(Y_1 \geq -y)) = \mathbb{P}(Y_1 \leq y) \end{aligned}$$

So Y_2 is normally distributed. However, $Y_1 + Y_2$ is normally distributed - consider the following

$$\begin{aligned} F_{Y_1+Y_2}(y) &= \mathbb{P}(Y_1 + Y_2 \leq y) \\ &= \mathbb{P}(Y_1 + Y_2 \leq y \,|\, Z=1)\mathbb{P}(Z=1) + \mathbb{P}(Y_1 + Y_2 \leq y \,|\, Z=-1)\mathbb{P}(Z=-1) \\ &= \frac{1}{2}(\mathbb{I}_{\{y \geq 0\}}(y) + \mathbb{P}(Y_1 \leq y/2)) \end{aligned}$$

so is not equal to CDF of a normal distribution (notice this is discontinuous at $y=0$). Therefore $\vec{Y}$ is not multivariate normally distributed in our case. The following is a simulation of this special distribution:

```
n = 1000
Y1 = randn(n)
Z = rand([-1,1],n)
Y2 = Y1.*Z
```

This time we plot a density estimate (or histogram) and compare with the actual density. You will see they are quite close.

```
using StatsPlots
# You may also use histogram.
pltAI = density(Y1,label="experimental",title="Y1");
plot!(x->pdf(Normal(),x),-4,4,label="actual")
pltAII = density(Y2,label="experimental",title="Y2");
plot!(x->pdf(Normal(),x),-4,4,label="actual")
pltA = plot(pltAI,pltAII,layout=(1,2),size=(800,300))
```

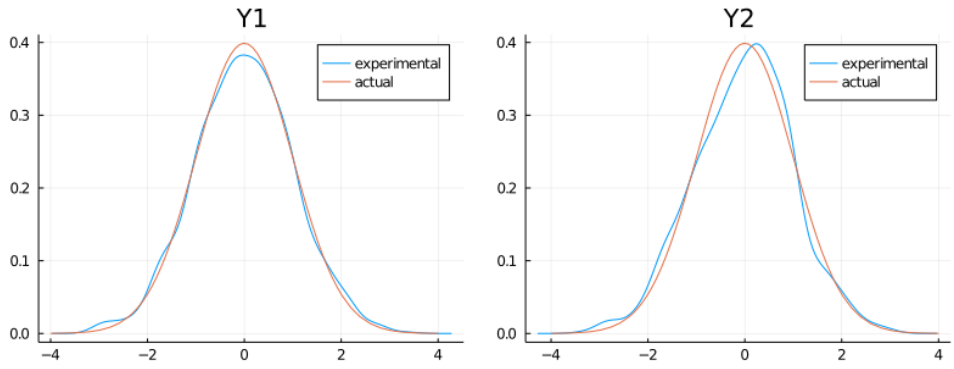

Figure 1.4: Densities of Y_1 and Y_2

But $Y_1 + Y_2$ is definitely not normally distributed. You will see a spike at $y = 0$. If we plot a scatter plot of Y_2 against Y_1 you will see a cross instead of our familiar ellipsoidal region.

```
pltBI = histogram(Y1.+Y2,nbins=80,...)
pltBII = scatter(Y1,Y2,...)
pltB = plot(pltBI,pltBII,layout=(1,2),size=(800,300))
```

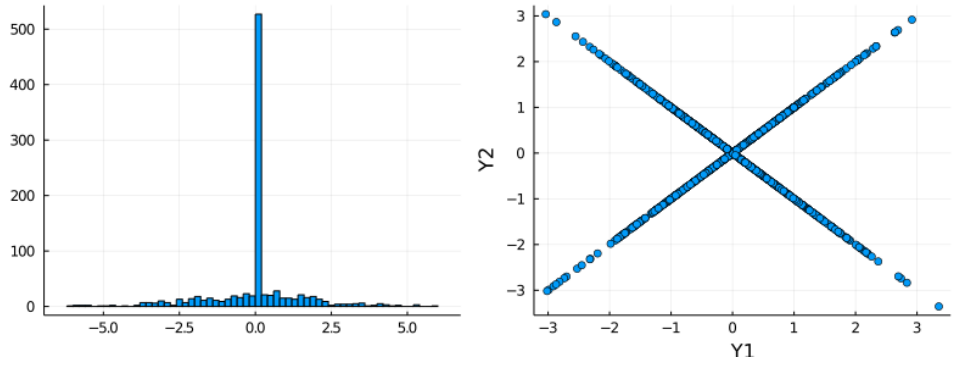

Figure 1.5: (Left) Density of $Y_1 + Y_2$. (Right) Y_2 against Y_1 scatter plot.

Similar to the (univariate) normal distribution, an affine transformation of MVN distributed random vector is also MVN distributed.

Exercise 1.3.7

Assume $\vec{X} \in \mathbb{R}^n$ is MVN distributed with mean $\vec{\mu}$ and covariance matrix Σ. Let $A \in \mathbb{R}^{m\times n}$ and $\vec{c} \in \mathbb{R}^m$. Show that $A\vec{X} + \vec{c}$ is normally distributed. What are the new mean and variance.

Hint. Check definition to show that $A\vec{X} + \vec{c}$ is MVN distributed. From Exercise 1.3.5 we can immediately obtain $A\vec{\mu} + \vec{c}$ as new mean and $A\Sigma A^T$ as new covariance.

This has given us a way to construct a MVN distributed random vector $\vec{X}$ which has mean vector $\vec{\mu}$ and covariance matrix Σ. As we have established in Exercise 1.3.5 that any covariance matrix Σ is symmetric and non-negative definite, we know that Σ has a Cholesky Decomposition $\Sigma = LL^T$ with $L \in \mathbb{R}^{n\times n}$ [7] (in particular if Σ is positive definite then L is full rank (hence invertible)). From the exercise above, the random vector $\vec{X} := L\vec{Z} + \vec{\mu}$ with $\vec{Z}$ vector of standard Gaussians is MVN distributed with required mean and variance.

Recall that we can characterise a normal distribution with its mean and variance. We will show that we can characterise an MVN with its mean and variance in general. The first attempt will be to look at its density. Notice the density [8] of vector

[7] There will be more discussion on Chapter 2

[8] $\|\vec{z}\|$ is the Euclidean norm if not specified otherwise

of standard Gaussian is

$$f_{\vec{Z}}(\vec{z}) = \frac{1}{(2\pi)^{n/2}} e^{-\|z\|^2/2} = \frac{1}{(2\pi)^{n/2}} e^{-(z_1^2+...+z_n^2)/2}, \ \vec{z} = (z_1, ..., z_n) \tag{1.14}$$

If $\vec{X}$ is MVN distributed with mean $\vec{\mu}$ and **positive definite** covariance matrix $\Sigma = LL^T$ (with L invertible), then we know that it has the same distribution as $g(\vec{Z})$, $g(\vec{z}) = L\vec{z} + \vec{\mu}$. The transformation is invertible with $g^{-1}(\vec{x}) = L^{-1}(\vec{x} - \vec{\mu})$. Therefore we can use the following theorem

Theorem 1.3.8: Continuous Multivariate Transformation Theorem

Let $\vec{X}$ be a random vector taking value in open set $A \subseteq \mathbb{R}^n$ and has density $f_{\vec{X}}(\vec{x})$. Let $g : A \to B$, $B \subseteq \mathbb{R}^n$ be a diffeomorphism (differentiable bijective mapping with differentiable inverse). Then the random variable $\vec{Y}$ has density

$$f_{\vec{Y}}(\vec{y}) = f_{\vec{X}}(g^{-1}(\vec{y})) \, |\det(Df^{-1}(\vec{y}))| \, \mathbb{I}_B(\vec{y}) \tag{1.15}$$

where $Df^{-1}(\vec{y})$ is the Jacobian (total derivative) of f^{-1}.

Remark. If $\vec{f} := (f_1, ..., f_n)$ is a differentiable function then its Jacobian $D\vec{f}(\vec{x}) = \dfrac{\partial \vec{f}}{\partial \vec{x}}$ is defined componentwise as $[D\vec{f}(x)]_{ij} = \dfrac{\partial f_i}{\partial x_j}$, or as followed:

$$D\vec{f}(\vec{x}) = \begin{pmatrix} \uparrow & \uparrow & & \uparrow \\ \frac{\partial \vec{f}}{\partial x_1} & \frac{\partial \vec{f}}{\partial x_2} & ... & \frac{\partial \vec{f}}{\partial x_n} \\ \downarrow & \downarrow & & \downarrow \end{pmatrix} \tag{1.16}$$

to conclude that $\vec{X}$ has the following density

Proposition 1.3.9: Density of MVN distributed vector

$$f_{\vec{X}}(\vec{x}) = \frac{1}{(2\pi)^{n/2}(\det(LL^T))^{1/2}} \exp\left(-\frac{(\vec{x} - \vec{\mu})^T (LL^T)^{-1} (\vec{x} - \vec{\mu})}{2}\right) \tag{1.17}$$

Proof.

$$\begin{aligned} f_{\vec{X}}(\vec{x}) &= f_{\vec{z}}(L^{-1}(\vec{x} - \vec{\mu}))|\det(L^{-1})| \\ &= \frac{1}{(2\pi)^{n/2}\det(L)} \exp\left(-\frac{1}{2}\left(L^{-1}(\vec{x} - \vec{\mu})\right)^T (L^{-1}(\vec{x} - \vec{\mu}))\right) \\ &= \frac{1}{(2\pi)^{n/2}(\det(LL^T))^{1/2}} \exp\left(-\frac{(\vec{x} - \vec{\mu})^T (LL^T)^{-1} (\vec{x} - \vec{\mu})}{2}\right) \\ &= \frac{1}{(2\pi)^{n/2}(\det(\Sigma))^{1/2}} \exp\left(-\frac{(\vec{x} - \vec{\mu})^T \Sigma^{-1} (\vec{x} - \vec{\mu})}{2}\right) \end{aligned}$$

□

Notice the density is solely determined by $\vec{\mu}$ and Σ, and hence MVN is characterised by its mean and variance. There is a loophole in the proof, though – when Σ is not positive definite then its inverse (hence density in B) does not exist.

> *Remark.* In general $\vec{X}$ has density in the column span of Σ, or $\text{Csp}(\Sigma)$.

To establish the fact that MVN is characterised by $\vec{\mu}$ and $\vec{\Sigma}$, we need to make use of *Moment Generating Function* (MGF) or *Characteristic Function* (CF). Recall if X is a univariate random variable then its MGF and CF are $M_X(t) = \mathbb{E}(e^{tX})$ and $\phi_X(t) = \mathbb{E}(e^{itX})$ respectively.

> *Remark.* They corresponds to Laplace Transform and Fourier Transform of density of X respectively, if the density exists. Also for random variables X, Y we have $\mathbb{E}(X + iY) = \mathbb{E}(X) + i\mathbb{E}(Y)$

We may generalise the idea to random vectors

Definition 1.3.10: MGF and CF of random vectors

If $\vec{X}$ is a random vector then the MGF and CF are defined as

$$M_{\vec{X}}(\vec{t}) = \mathbb{E}(e^{\vec{t}^T\vec{X}}) \tag{1.18}$$

$$\phi_{\vec{X}}(\vec{t}) = \mathbb{E}(e^{i\vec{t}^T\vec{X}}) \tag{1.19}$$

They have some properties in common (provided they **both** exists):

- $M_{\vec{X}}(\vec{0}) = \phi_{\vec{X}}(\vec{0}) = 1$
- (Characterisation) Random vectors $\vec{X}$ has same distribution as $\vec{Y} \iff M_{\vec{X}}(\vec{t}) = M_{\vec{Y}}(\vec{t}) \iff \phi_{\vec{X}}(\vec{t}) = \phi_{\vec{Y}}(\vec{t})$
- (Continuity) Random Vectors $(\vec{X}_i)_{i\geq 1}$ converges pointwise to random vector $\vec{X} \iff M_{\vec{X}_i}(\vec{t}) \to M_{\vec{X}}(\vec{t})$ pointwise $\iff \phi_{\vec{X}_i}(\vec{t}) \to \phi_{\vec{X}}(\vec{t})$ pointwise.
- (Independence) Random Vectors $\vec{X}$, $\vec{Y}$ are independent $\iff M_{(\vec{X},\vec{Y})(\vec{t})} = M_{\vec{X}}(\vec{t})M_{\vec{Y}}(\vec{t}) \iff \phi_{(\vec{X},\vec{Y})(\vec{t})} = \phi_{\vec{X}}(\vec{t})\phi_{\vec{Y}}(\vec{t})$
- (Affine Transformation in 1D) If X is 1D then $M_{\mu+\sigma X} = e^{\mu t}M_X(\sigma t)$ and $\phi_{\mu+\sigma X} = e^{i\mu t}\phi_X(\sigma t)$

But we would prefer using CF over MGF, because MGF does not always exist (consider Cauchy distribution with undefined mean) but CF does (and is always uniformly continuous). Moreover we know that the density of standard normal distribution is invariant over Fourier transform (i.e. if $X \sim \mathbf{N}(0,1)$ then $\phi_X(t) = e^{-t^2/2}$), but is not invariant over Laplace transform. In fact,

Proposition 1.3.11: CF of MVN Distributed Random Vector

If $\vec{X}$ is MVN distributed with mean $\vec{\mu}$ and covariance matrix Σ then its Characteristic Function is

$$\phi_{\vec{X}}(\vec{t}) = \exp\left(i\vec{t}^T\vec{\mu} - \frac{1}{2}\vec{t}^T\Sigma\vec{t}\right) \tag{1.20}$$

and therefore an MVN is characterised by its mean and covariance.

Proof. Consider the random variable $Y_{\vec{t}} = \vec{t}^T\vec{X}$. Then it is $(\mathbf{N}(\vec{t}^T\vec{\mu}, \vec{t}^T\Sigma\vec{t}))$ distributed. Therefore the characteristic function of $\vec{Y_t}$ is

$$\phi_{Y_t}(s) = \exp\left(i(\vec{t}^T\vec{\mu})s - \frac{s^2}{2}\vec{t}^T\Sigma\vec{t}\right) \tag{1.21}$$

The result hence follow by noting that $\phi_{\vec{X}}(\vec{t}) = \phi_{Y_{\vec{t}}}(1)$. which is dependent of $\vec{\mu}$ and Σ only. Hence (by characterisation) an MVN distribution is characterised by its mean and variance. □

Now it makes sense to write

$$\vec{X} \sim \mathbf{N}_n(\vec{\mu}, \Sigma) \tag{1.22}$$

referring to the fact that $\vec{X} \in \mathbb{R}^n$ is MVN distributed with mean $\vec{\mu} \in \mathbb{R}^n$ and covariance $\Sigma \in \mathbb{R}^{n\times n}$. Here is another application of Proposition 1.3.11 -

Exercise 1.3.12: Rotational Invariance of Vector of Standard Gaussians

If $\vec{Z}$ is a vector of standard Gaussian and Q is an orthogonal matrix then $Q\vec{Z}$ has same distribution as $\vec{Z}$

Hint. Look at the mean and covariance of $Q\vec{Z}$

The MVN distribution has two important (and intriguing) properties: (1) Uncorrelated $\iff$ Independence and (2) Central Limit Theorem.

Proposition 1.3.13: Independence of MVN Distributed Vectors

Let $\vec{X}_1 \sim \mathbf{N}_m(\vec{\mu}_1, \Sigma_{11})$ and $\vec{X}_2 \sim \mathbf{N}_n(\vec{\mu}_2, \Sigma_{22})$. Then $\vec{X}_1, \vec{X}_2$ are independent $\iff$ they are uncorrelated ($\Sigma_{12} := \text{Cov}(\vec{X}, \vec{Y}) = 0$)

Proof. The $\iff$ side is trivial (look at individual entries). For the other part, notice if $\vec{t} = (\vec{t}_1, \vec{t}_2)$ with $\vec{t}_1 \in \mathbb{R}^m$ and $\vec{t}_2 \in \mathbb{R}^n$ then

$$\begin{aligned}\phi_{(\vec{X}_1,\vec{X}_2)}(\vec{t}) &= \exp\left(i\begin{pmatrix}\vec{t}_1^T & \vec{t}_2^T\end{pmatrix}\begin{pmatrix}\vec{\mu}_1\\ \vec{\mu}_2\end{pmatrix} - \frac{1}{2}\begin{pmatrix}\vec{t}_1^T & \vec{t}_2^T\end{pmatrix}\begin{pmatrix}\Sigma_{11} & 0\\ 0^T & \Sigma_{22}\end{pmatrix}\begin{pmatrix}\vec{t}_1\\ \vec{t}_2\end{pmatrix}\right)\\ &= \exp\left(i\left(\vec{t}_1^T\vec{\mu}_1 + \vec{t}_2^T\vec{\mu}_2\right) - \frac{1}{2}\left(\vec{t}_1^T\Sigma_{11}\vec{t}_1 + \vec{t}_2^T\Sigma_{22}\vec{t}_2\right)\right) = \phi_{\vec{X}_1}(\vec{t}_1)\phi_{\vec{X}_2}(\vec{t}_2)\end{aligned}$$

hence result. □

Remark. It is also easy to see from this proof that the marginal distribution of an MVN is also MVN distributed. In particular, if we can partition $\vec{X} := \begin{pmatrix}\vec{X}_1\\ \vec{X}_2\end{pmatrix} \sim \mathbf{N}_n\left(\begin{pmatrix}\vec{\mu}_1\\ \vec{\mu}_2\end{pmatrix}, \begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{12}^T & \Sigma_{22}\end{pmatrix}\right)$ with $\vec{X}_1, \vec{\mu}_1 \in \mathbb{R}^k$, $\vec{X}_2, \vec{\mu}_2 \in \mathbb{R}^{n-k}$, $\Sigma_{11} \in \mathbb{R}^{k\times k}$, $\Sigma_{22} \in \mathbb{R}^{(n-k)\times(n-k)}$ then the marginals satisfies $\vec{X}_1 \sim \mathbf{N}_k(\vec{\mu}_1, \Sigma_{11})$ and $\vec{X}_2 \sim \mathbf{N}_{n-k}(\vec{\mu}_2, \Sigma_{22})$. (Just set $\vec{t}_1$ or $\vec{t}_2$ to zero vector)

An important corollary is the conditional distribution of MVN vectors:

Corollary 1.3.14: Conditional Distribution

Again if $\vec{X} := \begin{pmatrix} \vec{X}_1 \\ \vec{X}_2 \end{pmatrix} \sim \mathbf{N}_n \left(\begin{pmatrix} \vec{\mu}_1 \\ \vec{\mu}_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{pmatrix} \right)$ as above then

$$\vec{X}_1 \,|\, \vec{X}_2 \sim \mathbf{N}_k \left(\vec{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\vec{X}_2 - \vec{\mu}_2), \Sigma_{11.2} \right)$$

where $\Sigma_{11.2} := \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T$.

Proof. Consider the matrix

$$A = \begin{pmatrix} I^{(k)} & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0^T & I^{(n-k)} \end{pmatrix} \tag{1.23}$$

We then have

$$A\vec{X} = \begin{pmatrix} \vec{X}_1 - \Sigma_{12}\Sigma_{22}^{-1}\vec{X}_2 \\ \vec{X}_2 \end{pmatrix} \sim \mathbf{N}_n \left(A \begin{pmatrix} \vec{\mu}_1 \\ \vec{\mu}_2 \end{pmatrix}, A \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{pmatrix} A^T \right)$$

Notice that

$$\begin{aligned} A \begin{pmatrix} \vec{\mu}_1 \\ \vec{\mu}_2 \end{pmatrix} &= \begin{pmatrix} \vec{\mu}_1 - \Sigma_{12}\Sigma_{22}^{-1}\vec{\mu}_2 \\ \vec{\mu}_2 \end{pmatrix} \\ A \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{pmatrix} A^T &= \begin{pmatrix} I^{(k)} & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0^T & I^{(n-k)} \end{pmatrix} \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{pmatrix} \begin{pmatrix} I^{(k)} & 0 \\ -(\Sigma_{22}^{-1})^T\Sigma_{12}^T & I^{(n-k)} \end{pmatrix} \\ &= \begin{pmatrix} I^{(k)} & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0^T & I^{(n-k)} \end{pmatrix} \begin{pmatrix} \Sigma_{11} - \Sigma_{12}(\Sigma_{22})^{-1}\Sigma_{12}^T & \Sigma_{12} \\ 0^T & \Sigma_{22} \end{pmatrix} \\ &= \begin{pmatrix} \Sigma_{11.2} & 0 \\ 0^T & \Sigma_{22} \end{pmatrix} \end{aligned}$$

So we have

$$\begin{pmatrix} \vec{X}_1 - \Sigma_{12}\Sigma_{22}^{-1}\vec{X}_2 \\ \vec{X}_2 \end{pmatrix} \sim \mathbf{N}_n \left(\begin{pmatrix} \vec{\mu}_1 - \Sigma_{12}\Sigma_{22}^{-1}\vec{\mu}_2 \\ \vec{\mu}_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11.2} & 0 \\ 0^T & \Sigma_{22} \end{pmatrix} \right)$$

In particular $\vec{X}_1 - \Sigma_{12}\Sigma_{22}^{-1}\vec{X}_2$ is independent with $\vec{X}_2$. Therefore

$$\vec{X}_1 - \Sigma_{12}\Sigma_{22}^{-1}\vec{X}_2 \,|\, \vec{X}_2 \sim \mathbf{N}_k \left(\vec{\mu}_1 - \Sigma_{12}\Sigma_{22}^{-1}\vec{\mu}_2,\ \Sigma_{11.2} \right)$$

and hence result, i.e. $\vec{X}_1 \,|\, \vec{X}_2 \sim \mathbf{N}_k \left(\vec{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\vec{X}_2 - \vec{\mu}_2), \Sigma_{11.2} \right)$. Notice how this formula resembles (1.4). □

We also establish the Central Limit Theorem for random vectors.

Theorem 1.3.15: Multivariate Central Limit Theorem (CLT)

Let $\vec{X}_1, \vec{X}_2, ... \in \mathbb{R}^n$ be iid sequence of random vectors with mean $\vec{\mu}$ and variance $\vec{\Sigma}$. Let $\bar{X}_N := N^{-1}\sum_{i=1}^N \vec{X}_i$ be the sample mean. Then as $N \to \infty$, $N^{1/2}(\vec{X}_N - \vec{\mu})$ converges in distribution to a $\mathbf{N}_n(\vec{0}, \Sigma)$ distributed vector.

Proof. We let (with $\vec{t} \in \mathbb{R}^n$ fixed)

$$\vec{Y}^{(N)} = N^{1/2}(\bar{X}_N - \vec{\mu}) = N^{-1/2}\sum_{i=1}^N (\vec{X}_i - \vec{\mu}) \tag{1.24}$$

$$Y_{\vec{t}}^{(N)} = \vec{t}^T\vec{Y}^{(N)} = N^{-1/2}\sum_{i=1}^N (\vec{t}^T\vec{X}_i - \vec{t}^T\vec{\mu}) \tag{1.25}$$

Notice that the sequence $(\vec{t}^T\vec{X}_i - \vec{t}^T\vec{\mu})_{i\geq 1}$ is iid with mean 0 and covariance $\vec{t}^T\Sigma\vec{t}$. Hence $Y_{\vec{t}}^{(N)}$ converges in distribution to $\mathbf{N}(0, \vec{t}^T\Sigma\vec{t})$ by (univariate) CLT. It follows that for all $\vec{t}$, the CF of $\vec{Y}_{\vec{t}}^{(N)}$ converges pointwise:

$$\forall s \in \mathbb{R}, \quad \phi_{Y_{\vec{t}}^{(N)}}(s) \xrightarrow{\mathcal{D}} \exp\left(-\frac{s^2}{2}(\vec{t}^T\Sigma\vec{t})\right)$$

Plugging $s = 1$ we have

$$\phi_{\vec{Y}^{(N)}}(\vec{t}) = \phi_{Y_{\vec{t}}^{(N)}}(s) \xrightarrow{\mathcal{D}} \exp\left(-\frac{1^2}{2}(\vec{t}^T\Sigma\vec{t})\right)$$

The result follows by continuity of CF. □

So the multivariate normal distribution is phenomenal.

1.3.3 (Shifted) Chi-Squared $\chi_n^2(\delta)$ Distributions

We define the distribution for quadratic forms.

Definition 1.3.16: $\chi_n^2(\delta)$ Distribution

Let $\vec{Z} = (Z_1, ..., Z_n) \sim \mathbf{N}_n(\vec{\mu}, I^{(n)})$ with $\vec{\mu} \in \mathbb{R}^n$. Then the (non-central) χ^2 distribution (with degree of freedom n) and dispersion factor $\delta = \|\vec{\mu}\|^2$ is defined as the distribution of $\left\|\vec{Z}\right\|^2 = \vec{Z}^T\vec{Z} = Z_1^2 + ... + Z_n^2$. In particular if $\vec{\mu} = \vec{0}$ then the distribution of $\vec{Z}^T\vec{Z}$ is known as the (central) χ^2 distribution (with degree of freedom n), or $\chi_n^2 := \chi_n^2(0)$.

Remark. Notice that it depends with the norm $\vec{Z}$ only but not any individual entries of $\vec{Z}$ – notice that if $\vec{Z}' = Q\vec{Z}$ with Q orthogonal then $\vec{Z}'^T\vec{Z}' = \vec{Z}^TQ^TQ\vec{Z} = \vec{Z}^T\vec{Z}$. Indeed we can choose Q such that $Q\vec{Z} = (\|\vec{\mu}\|, 0, ..., 0)$. So $\vec{Z}^T\vec{Z} = \vec{Z}'^T\vec{Z}' = \chi_1^2(\|\vec{\mu}\|^2) + \chi_{n-1}^2$.

Notice that if $\vec{X} \sim \mathbf{N}_n(\vec{\mu}, \Sigma)$ with $\Sigma = LL^T$ for some L. If Σ is positively definite then L is always invertible, and therefore we have $\vec{Z} = L^{-1}(\vec{X} - \vec{\mu}) \sim \mathbf{N}_n(\vec{0}, I^{(n)})$. It follows that

$$(\vec{X} - \vec{\mu})^T \Sigma^{-1} (\vec{X} - \vec{\mu}) = \vec{Z}^T \vec{Z} \sim \chi^2_n \tag{1.26}$$

and similarly

$$\vec{X}^T \vec{X} \sim \chi^2_n(\delta), \quad \delta = \left\| L^{-1}\vec{\mu} \right\|^2 = \vec{\mu}^T \Sigma^{-1} \vec{\mu} \tag{1.27}$$

What do we know about this distribution? We can start by obtaining density of $\chi^2_1(\delta)$ by inversion:

Exercise 1.3.17: Density of $\chi^2_1(\delta)$

Recall that if $X \sim \mathbf{N}(\sqrt{\delta}, 1)$, then $Y = X^2 \sim \chi^2_1(\delta)$ Use this fact to obtain the density of Y.

Hint. Notice that for all $y \geq 0$, the CDF of Y is

$$F_Y(y) = \mathbb{P}(-\sqrt{y} \leq X \leq \sqrt{y}) = \frac{1}{\sqrt{2\pi}} \int_{-\sqrt{y}}^{\sqrt{y}} \exp\left(-\frac{1}{2}(t - \sqrt{\delta})^2 \right) \tag{1.28}$$

By differentiation we have

$$\begin{aligned} f_Y(y) &= \frac{1}{\sqrt{2\pi}} \frac{d}{dy} \left(\int_{-\sqrt{y}}^{\sqrt{y}} \exp\left(-\frac{1}{2}(t - \sqrt{\delta})^2 \right) \right) \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{2\sqrt{y}} \left(\exp\left(-\frac{1}{2}(\sqrt{y} - \sqrt{\delta})^2 \right) + \exp\left(-\frac{1}{2}(-\sqrt{y} - \sqrt{\delta})^2 \right) \right) \\ &= \frac{1}{\sqrt{2\pi y}} \exp\left(-\frac{y + \delta}{2} \right) \frac{1}{2} \left(e^{-\sqrt{\delta y}} + e^{\sqrt{\delta y}} \right) \\ &= \frac{1}{\sqrt{2\pi y}} \exp\left(-\frac{y + \delta}{2} \right) \cosh\left(\sqrt{\delta y} \right) \end{aligned}$$

When $\delta = 0$, the density simplifies to

$$f_Y(y) = \frac{1}{\sqrt{2\pi}} y^{-1/2} e^{-y/2} \tag{1.29}$$

So Y is $\Gamma(1/2, 1/2)$ [a] distributed in this case.

[a] We adopt the shape-scale parametrisation, that if Y is $\Gamma(\alpha, \beta)$ distributed iff Y has density $f_Y(y) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$, with $\Gamma(\alpha)$ the Gamma distribution.

Remark. Sometimes we would expand $\cosh\left(\sqrt{\delta y}\right)$ in the above expression. We will see why in latter argument.

$$f_Y(y) = \frac{1}{\sqrt{2\pi}} e^{-\delta/2} \sum_{i=0}^{\infty} \frac{\delta^i}{(2i)!} y^{i - \frac{1}{2}} e^{-\frac{y}{2}} \tag{1.30}$$

But the density above is difficult to be dealt with – especially when it comes to convolution. We turn to studying the CF of $\chi^2_n(\delta)$.

Proposition 1.3.18: CF of $\chi^2_n(\delta)$

Let $\vec{Z} = (Z_1, ..., Z_n) \sim \mathbf{N}_n(\vec{\mu}, I^{(m)})$ with $\vec{\mu} \in \mathbb{R}^m$ and $\delta = \vec{\mu}^T\vec{\mu}$. Let $Y = \vec{Z}^T\vec{Z}$. Then Y has characteristic function

$$\phi_Y(t) = \mathbb{E}(e^{itY}) = (1-2it)^{-n/2} \exp\left(\frac{i\delta t}{1-2it}\right) \tag{1.31}$$

Proof. I wonder are you dare to directly perform Fourier transform of the density in (1.29). There is a much smarter way:

$$\begin{aligned}\phi_Y(t) = \mathbb{E}(e^{it\vec{Z}^T\vec{Z}}) &= \int_{\mathbb{R}^n} \frac{1}{(2\pi)^{n/2}} \exp\left(it\vec{z}^T\vec{z}\right) \exp\left(-\frac{1}{2}(\vec{z}-\vec{\mu})^T(\vec{z}-\vec{\mu})\right) d\vec{z} \\ &= \int_{\mathbb{R}^n} \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\left((1-2it)\vec{z}^T\vec{z} - 2\vec{\mu}^T\vec{z} + \vec{\mu}^T\vec{\mu}\right)\right) d\vec{z}\end{aligned}$$

Notice that

$$\begin{aligned}&(1-2it)\vec{z}^T\vec{z} - 2\vec{\mu}^T\vec{z} + \vec{\mu}^T\vec{\mu} \\ &= (1-2it)\left(\vec{z}^T\vec{z} - \frac{2}{1-2it}\vec{\mu}^T\vec{z} + \frac{\vec{\mu}^T\vec{\mu}}{(1-2it)^2}\right) + \delta - \frac{\vec{\mu}^T\vec{\mu}}{1-2it} \\ &= (1-2it)\left(\vec{z}-\vec{\mu}'\right)^T\left(\vec{z}-\vec{\mu}'\right) - \frac{2it\delta}{1-2it}, \quad \vec{\mu}' = \frac{1}{1-2it}\vec{\mu}\end{aligned}$$

Therefore

$$\begin{aligned}\phi_Y(t) &= \int_{\mathbb{R}^n} \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2(1-2it)^{-1}}\left(\left(\vec{z}-\vec{\mu}'\right)^T\left(\vec{z}-\vec{\mu}'\right)\right)\right) \exp\left(\frac{it\delta}{1-2it}\right) d\vec{z} \\ &= (1-2it)^{-n/2} \exp\left(\frac{it\delta}{1-2it}\right)\end{aligned}$$

□

We can therefore establish the following important results for $Y \sim \chi^2_n(\delta)$:

Exercise 1.3.19: Important Results of $\chi^2_n(\delta)$

1. Y has same distribution as the sum of $Y_1 \sim \chi^2_1(\delta)$ and $Y_2, ..., Y_n \overset{\text{iid}}{\sim} \chi^2_1$. In particular when $\delta = 0$ then Y has same distribution as n iid χ^2_1-distributed variables.
2. $\mathbb{E}(Y) = n + \delta$ and $\mathrm{Var}(Y) = 2n + 4\delta$.
3. If $Y_1 \sim \chi^2_{n_1}(\delta_1)$ and $Y_2 \sim \chi^2_{n_1}(\delta_2)$ then $Y_1 + Y_2 \sim \chi^2_{n_1+n_2}(\delta_1+\delta_2)$.

Hint. Decompose $\phi_Y(t)$ into suitable products yields (1). Differentiating $\phi_Y(t)$ would yield (2). (3) follows by multiplying the CFs.

Remark. By addition rule of Gamma distribution we know that $U \sim \chi^2_n(\delta)$ has the same distribution as $\Gamma(n/2, 1/2)$. Therefore U has density

$$f_U(u) = \frac{1}{2^{n/2}\Gamma(n/2)} u^{k/2-1} e^{-u/2} \tag{1.32}$$

Let us study equation (1.30) again. Notice the following identity hold: [a]

$$\begin{aligned}\Gamma\left(i+\frac{1}{2}\right) &= \left((i-1)+\frac{1}{2}\right)\left((i-2)+\frac{1}{2}\right)...\left(\frac{1}{2}\right)\Gamma\left(\frac{1}{2}\right) \\ &= \frac{(2i-1)\times(2i-3)...5\times 3\times 1}{2^i}\sqrt{\pi} = \frac{(2i)!}{2^{2i}i!}\sqrt{\pi}\end{aligned}$$

Therefore if $Y \sim \chi^2_1(\delta)$

$$\begin{aligned}f_Y(y) &= \frac{1}{\sqrt{2\pi}} e^{-\delta/2} \sum_{i=0}^{\infty} \frac{\delta^i}{(2i)!} y^{i-\frac{1}{2}} e^{-\frac{y}{2}} \\ &= \frac{e^{-\delta/2}}{\sqrt{2}} \sum_{i=0}^{\infty} \frac{\delta^i}{2^{2i} i! \Gamma(i+1/2)} y^{\frac{2i+1}{2}-1} e^{-\frac{y}{2}} \\ &= e^{-\delta/2} \sum_{i=0}^{\infty} \frac{1}{i!}\left(\frac{\delta}{2}\right)^i \left(\frac{1}{2^{\frac{2i+1}{2}}\Gamma\left(\frac{2i+1}{2}\right)} y^{\frac{2i+1}{2}-1} e^{-\frac{y}{2}}\right)\end{aligned}$$

In other words, $\chi^2_1(\delta) \,|\, K \sim \chi^2_{1+2K},\ K \sim \text{Po}\,(\delta/2)$. Notice that $\chi^2_n(\delta) \sim \chi^2_1(\delta) + \chi^2_{n-1}$, therefore we have

$$\chi^2_n(\delta) \,|\, K \sim \chi^2_{n+2K}, \qquad K \sim \text{Po}\,(\delta/2) \tag{1.33}$$

Of course, we may write the density in terms of Hypogeometric Functions [b], but we will use this representation later. Perhaps relation (1.33) provides you a simpler way to obtain moments of $\chi^2_n(\delta)$ distribution.

[a] The following is the proof for the case when $i \geq 1$, but it is also true for $i = 0$.
[b] If you know what it means

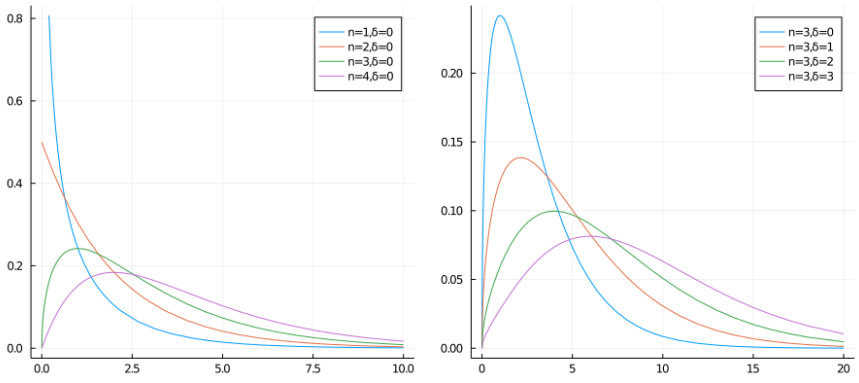

Figure 1.6: Densities of $\chi^2_n(\delta)$ for varying n and δ

1.3.4 Student-t and (Shifted) F-Distribution

Definition 1.3.20: Student-t Distributions

Let X be a random variable. X is said to follow t_n distribution (with degree of freedom n) if it has same distribution functions as $Z/(\sqrt{S/n})$, with Z, S independent, $Z \sim \mathbf{N}(0,1)$ and $S \sim \chi^2_n$.

One can see student-t distribution as a rescaled version of Normal distribution. In fact when $n \to \infty$ we have $X \to \mathbf{N}(0,1)$ in distribution. To prove this, notice S is a sum of n iid χ^2_1 distributed variable so by Weak Law of Large Number (WLLN) $S/n \to 1$ in probability. By Continuous Mapping Theorem (CMT), $1/\sqrt{S/n} \to 1$ in probability, so by Slutsky's Theorem $X \to Z$ in distribution.

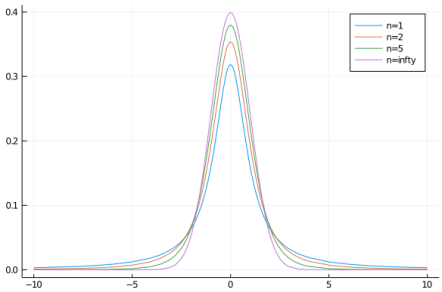

Figure 1.7: Density of t_n distribution

The use of this distribution comes from a corollary of Cochran's Theorem as followed:

Theorem 1.3.21: Sample Mean and Variance

Let $X_1, ..., X_n \overset{iid}{\sim} \mathbf{N}(\mu, \sigma^2)$. Define $\bar{X} = n^{-1}\sum_{i=1}^n X_i$ be the sample mean and $S^2 = (n-1)^{-1}\sum_{i=1}^n (X_i - \bar{X})^2$. Then

- $\bar{X}, S^2$ are independent.
- $\bar{X} \sim \mathbf{N}(\mu, \sigma^2/n)$
- $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$
- $T = \dfrac{\bar{X} - \mu}{\sqrt{S^2/n}} \sim t_{n-1}$

This theorem has been used to construct statistical tests for $\bar{X}$ and S^2, and we will generalise this theorem in Chapter 3 for linear models. Here we obtain the density of t-distribution

Exercise 1.3.22: Density of t-distribution

Show that if T follows a t_n distribution then it has density

$$f_T(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\Gamma\left(\frac{n}{2}\right)}\left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}} \tag{1.34}$$

Hint. Consider the transformation $g : (x, s) \mapsto (x/\sqrt{s/n}, s)$ on $\mathbb{R} \to \mathbb{R}_{>0}$. This is a diffeomorphism with $g^{-1} : (t, s) \to (t\sqrt{s/n}, s)$ and

$$Dg^{-1}(t,s) = \begin{pmatrix} \sqrt{s/n} & t/(2\sqrt{sn}) \\ 0 & 1 \end{pmatrix}, \quad \det(Dg^{-1}(t,s)) = \sqrt{\frac{s}{n}}$$

By Theorem 1.3.8, if $(Y, S) = g(X, S)$, then

$$\begin{aligned} f_{(T,S)}(t,s) &= f_X(t\sqrt{s/n}) f_S(s) \left|\det(Dg^{-1}(y,s))\right| \\ &= \left(\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2n}s\right)\right)\left(\frac{1}{2^{n/2}\Gamma(n/2)} s^{n/2-1} e^{-s/2}\right)\sqrt{\frac{s}{n}} \\ &= \frac{1}{\sqrt{2\pi n}} \frac{1}{2^{n/2}\Gamma(n/2)} s^{(n+1)/2-1} \exp\left(-\frac{1}{2}\left(\frac{t^2}{n}+1\right)s\right) \end{aligned}$$

Making the substitution $u = sc_n$ with $c_n = 1 + t^2/n$, we have

$$\begin{aligned} f_T(t) &= \frac{1}{\sqrt{2\pi n}} \frac{1}{2^{n/2}\Gamma(n/2)} \int_0^\infty (u/c_n)^{(k+1)/2-1} \exp(-u/2) \frac{du}{c} \\ &= \frac{c_n^{-\frac{n+1}{2}}}{\sqrt{2\pi n} 2^{n/2}\Gamma(n/2)} \int_0^\infty u^{(n+1)/2-1} e^{-u/2}\, du \\ &= \frac{2^{(n+1)/2}\Gamma(\frac{n+1}{2})}{\sqrt{2\pi n} 2^{n/2}\Gamma(n/2)} \left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}} \\ &= f_T(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\Gamma\left(\frac{n}{2}\right)} \left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}} \end{aligned}$$

Notice the special cases: when $\nu = 1$, T has a Cauchy(1) distribution. When $\nu \to \infty$ then T converges (distribution) to a standard normal $\mathbf{N}(0, 1)$ random variable.

Finally we state the definition of F distribution without further discussion.

Definition 1.3.23: Shifted F Distribution

Let X be a random variable. X is said to follow $F_{n_1,n_2}(\delta)$ distribution it has same distribution fuctions as $(X_1/n_1)/(X_2/n_2)$, where $X_1 \sim \chi^2_{n_1}(\delta)$ and $X_2 \sim \chi^2_{n_2}$.

The density can be obtained by considering similar transformation as in above exercise, which would involved calculation of Hypogeometric Functions, so is omitted at this stage.

CHAPTER 2

Ordinary Least Square

2.1 The Least Square Problem

The ordinary least square problem asks the following:

> *Question.* If $\vec{v} \in \mathbb{R}^n$ is a vector and $U = \text{Span}(\{\vec{u}_1, ..., \vec{u}_p\})$ is a subspace of $\mathbb{R}^n$ (with $\vec{u}_i \in \mathbb{R}^n$, $p \leq n$), can you find a vector $\vec{u}^* \in U$ which is a best approximation of $\vec{v}$.

Quantitatively, we want to minimise 2-norm of residual, i.e. find

$$\vec{u}^* = \underset{\vec{u} \in U}{\text{argmin}} \ \|\vec{v} - \vec{u}\|^2 \tag{2.1}$$

The vector $\vec{u}^*$ is known as a *projection* of $\vec{v}$ on U. Notice that $\forall \vec{u} \in U$, we may find unique constants $\beta_1, ...\beta_p \in \mathbb{R}$ such that $\vec{u} = \sum_{i=1}^{n} \beta_i \vec{u}_i$, therefore it is sufficient to find $\vec{\beta}^* = (\beta_1^*, ..., \beta_p^*)$ such that

$$\vec{\beta}^* = \underset{\vec{\beta} \in \mathbb{R}^p}{\text{argmin}} \ \left\|\vec{v} - X\vec{\beta}\right\|^2 \tag{2.2}$$

where X is the $n \times p$, matrix (with full rank (=p))

$$X = \begin{pmatrix} \uparrow & \uparrow & & \uparrow \\ \vec{u}_1 & \vec{u}_2 & ... & \vec{u}_p \\ \downarrow & \downarrow & & \downarrow \end{pmatrix} \tag{2.3}$$

Then we have $\vec{u}^* = X\vec{\beta}^*$. For convenience, define the *residual* (vector) of approximation be $\vec{e} = \vec{Y} - X\vec{\beta}$ with $f(\vec{\beta}) := \|\vec{e}\|^2 := \left\|\vec{v} - X\vec{\beta}\right\|^2$. Notice that

$$f(\vec{\beta}) = (\vec{v} - X\vec{\beta})^T(\vec{v} - X\vec{\beta}) = \vec{\beta}^T(X^TX)\vec{\beta} - 2\vec{v}^T(X\vec{\beta}) + \vec{v}^T\vec{v} \tag{2.4}$$

So it is a "quadratic function" of some kind. We will show later that $f(\vec{\beta})$ is a (globally) convex function (over $\mathbb{R}^p$) so it has a unique global minimum. The function $T' := T'_U$ sending $\vec{v}$ to $\vec{\beta}^*$ (and hence also the function $T := T_U$ sending $\vec{v}$ to $\vec{u}^*$).

2.1.1 Canonical Examples

To get more intuition, let consider two simple cases. The first case when $U = \text{Span}((1,1,...,1)^T) = \{\beta(1,1,...,1) : \beta \in \mathbb{R}\}$. If we write the i-th entry of $\vec{v}$ as v_i, then we have

$$\begin{aligned} f(\beta) = \sum_{i=1}^{n}(v_i - \beta)^2 &= n\beta^2 - 2\left(\sum_{i=1}^{n} v_i\right)\beta + \left(\sum_{i=1}^{n} v_i^2\right) \\ &= n\left(\beta - \frac{1}{n}\sum_{i=1}^{n} v_i\right)^2 + \left(\sum_{i=1}^{n} v_i^2\right) - \frac{1}{n}\left(\sum_{i=1}^{n} v_i\right)^2 \end{aligned}$$

So we immediate see that β^* (minimiser of $f(\beta)$) is the *sample mean* of entries of vector $\vec{v}$, or $\beta^* = n^{-1}\sum_{i=1}^{n} v_i$. Notice that

$$f(\beta^*) = n\left(\frac{1}{n}\left(\sum_{i=1}^{n} v_i^2\right) - \left(\frac{1}{n}\sum_{i=1}^{n} v_i\right)^2\right)$$

which is $(n-1)$ times the sample variance of entries of $\vec{v}$! So if $(v_1,...,v_n)$ happens to be the a sample from $(V_1,...,V_n)$ with $V_i = \beta + \epsilon_i$ (constant model, model 1 in Exercise 1.1.4) then

- the sample mean $\bar{V}$ is the "(ordinary) least square" (OLS) estimator of μ,
- the norm-squared of residual $f(\bar{V})$ divided by $n-1$ is an unbiased estimator of variance of V_i.

We may generalise a little bit and consider $U = \text{Span}(\vec{u}) = \{\beta\vec{u} : \beta \in \mathbb{R}\}$ with $\vec{u}$ non-zero vector in $\mathbb{R}^n$. Of course, if we write the i-th entry of $\vec{u}$ as u_i then we have

$$\begin{aligned} f(\beta) &= \sum_{i=1}^{n}(v_i - \beta u_i)^2 \\ &= \left(\sum_{i=1}^{n} u_i^2\right)\beta^2 - 2\left(\sum_{i=1}^{n} u_i v_i\right)\beta + \left(\sum_{i=1}^{n} v_i^2\right) \\ &= \left(\sum_{i=1}^{n} u_i^2\right)\left(\beta - \frac{\sum_{i=1}^{n} u_i v_i}{\sum_{i=1}^{n} u_i^2}\right)^2 + \left(\sum_{i=1}^{n} v_i^2\right) - \frac{\left(\sum_{i=1}^{n} u_i v_i\right)^2}{\left(\sum_{i=1}^{n} u_i\right)^2} \\ &= \|\vec{u}\|^2\left(\beta - \frac{\vec{u}^T\vec{v}}{\|\vec{u}\|^2}\right)^2 + \|\vec{v}\|^2 - \frac{(\vec{u}^T\vec{v})^2}{\|u\|^2} \end{aligned}$$

So $\beta^* = T'(\vec{v}) = (\vec{u}^T\vec{v})/\|u\|^2$ and

$$\vec{u}^* = T(\vec{v}) = \frac{(\vec{u}^T\vec{v})}{\|u\|^2}\vec{u} \tag{2.5}$$

> *Remark.* Note that $f(\beta) \geq 0$ for all β, in particular $f(\beta^*) \geq 0$. Rearranging yields *Cauchy-Schwarz* Inequality. Equality holds iff $\vec{v} \in \text{Span}(\vec{u})$.

But is there any geometric intuition behind the formula?

If we look carefully at the figure below, we must have residual $\vec{u}^{\perp} := \vec{v} - \vec{u}^{*}$ perpendicular to $\vec{u}^{*}$, otherwise the norm of residual is not minimised. This is why $\vec{u}^{*}$ is known as an orthogonal projection of $\vec{v}$ on U.

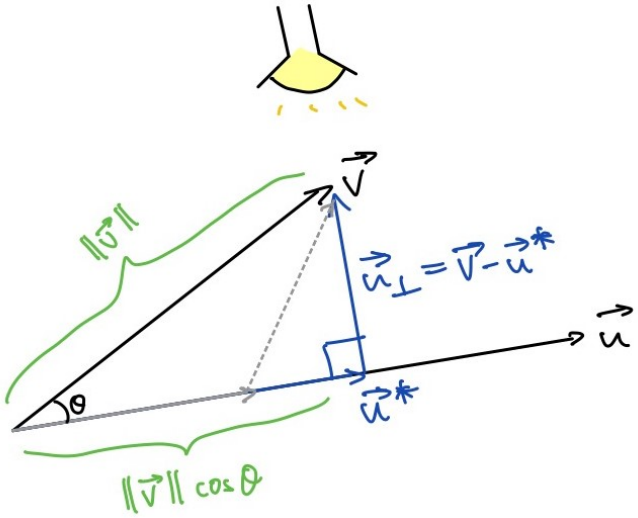

Figure 2.1: Vector Projection - you have seen this in high school

As a sanity check:

$$\vec{u}^{T}\vec{u}^{\perp} = \vec{u}^{T}\left(\vec{v} - \frac{(\vec{u}^{T}\vec{v})}{\|u\|^{2}}\vec{u}\right) = \vec{u}^{T}\vec{v} - \frac{(\vec{u}^{T}\vec{v})}{\|u\|^{2}}\|u\|^{2} = 0$$

Remark. The Cauchy-Schwarz Inequality enables us to define size of angle between two vectors. We assign the size of angle between $\vec{u}$ and $\vec{v}$ as the value θ:

$$\theta = \cos^{-1}\left(\frac{\vec{u}^{T}\vec{v}}{\|\vec{u}\|\,\|\vec{v}\|}\right) \tag{2.6}$$

If we relate back to the high-school geometry we have learnt, it is not hard to see that

$$\vec{u}^{*} = (\|\vec{v}\|\cos\theta)\frac{\vec{u}}{\|\vec{u}\|} = \left(\frac{\vec{u}^{T}\vec{v}}{\|\vec{u}\|^{2}}\right)\vec{u}$$

some even use this as an official derivation of the formula for projection of $\vec{v}$ on $\vec{u}$. Be careful that for the definition of angle between two vectors to make sense, we need *Cauchy-Schwarz Inequality*; and before we have this inequality, we already have a formula of projection.

Finally, we note from formula (2.5) that the map $T(\vec{v})$ is linear (notice the (bi-) linearity of inner (dot) product). We might therefore represent the projection as a matrix. Note that if $E = \{\vec{e}_i\}_{i=1}^{n}$ represents the standard basis of $\mathbb{R}^n$ (with

$\vec{e}_i = (0, ..., 0, 1, 0, ..., 0)$) we have

$$T_U(\vec{e}_i) = \left(\frac{u_i}{\|u\|^2}\right)\vec{u} = \frac{1}{(\vec{u}^T\vec{u})}\begin{pmatrix} u_1u_i \\ \cdots \\ u_iu_i \\ \cdots \\ u_nu_i \end{pmatrix}$$

Therefore we have $T_U(\vec{v}) = P\vec{v}$, where

$$P = \left(\vec{u}^T\vec{u}\right)^{-1}\begin{pmatrix} u_1u_1 & u_1u_2 & \cdots & u_1u_n \\ u_2u_1 & u_2u_2 & \cdots & u_2u_n \\ \cdots & \cdots & \cdots & \cdots \\ u_nu_1 & u_nu_2 & \cdots & u_nu_n \end{pmatrix} = \vec{u}\left(\vec{u}^T\vec{u}\right)^{-1}\vec{u}^T \tag{2.7}$$

Exercise 2.1.1

Is P symmetric? What rank does P have? What eigenvalues does P have? What algebraic multiplicities does the eigenvalues have?

Hint. P is symmetric. It has rank 1. Notice that $P^2 = P$ (why?), therefore if λ is an eigenvalue with (non-zero) eigenvector $\vec{x}$ we have $P^2\vec{x} = P\lambda\vec{x} = \lambda^2\vec{x}$. Therefore $\lambda^2 = \lambda \iff \lambda = 0, 1$. Finally, note that algebraic multiplicity of 1 (or $a(1)$) is in fact rank of matrix P (why?), so $a(1) = 1$ and $a(0) = n - 1$.

2.1.2 Intermediate Step: Classical Gram Schmidt

To summarise the above argument - you can decompose a vector $\vec{v} \in \mathbb{R}$ (uniquely) into a sum of vectors – $\vec{u}^* + \vec{u}^\perp$, with $\vec{u}^* \in \text{Span}(\vec{u})$ and $\vec{u}^\perp \perp \vec{u}$ (hence perpendicular to all vectors in $\text{Span}(\vec{u})$. Can we generalise this argument to any subspace $U \leq \mathbb{R}^n$, i.e. for $\vec{v} \in \mathbb{R}^d$ we can write it as $\vec{u}^* + \vec{u}^\perp$ with $\vec{u}^* \in U$ and $\vec{u}^\perp$ perpendicular to any vectors in U? If so, then $T_U(\vec{v})$ is probably $\vec{u}^*$ (we need to check this!). As a matter of notation, define:

Definition 2.1.2: Orthogonal Complement

Let $U \leq \mathbb{R}^n$. The Orthogonal Complement of U is

$$U^\perp := \left\{\vec{z} \in \mathbb{R}^n \,|\, \forall \vec{u} \in U, \vec{z}^T\vec{y} = 0\right\} \tag{2.8}$$

Here are some quick exercises to see if you understand the above definition.

Exercise 2.1.3

Show that (1) $U^\perp$ is a (linear) subspace in $\mathbb{R}^n$ and (2) $U \cap U^\perp = \left\{\vec{0}\right\}$

Hint. 1. Follow definitions!
2. $\vec{u} \in U \cap U^\perp \implies \vec{u}^T\vec{u} = 0 \iff \vec{u} = \vec{0}$.

With this notion, we can ask whether we can decompose $\vec{v}$ uniquely into $\vec{u}^* + \vec{u}^\perp$ with $\vec{u}^* \in U$ and $\vec{u}^\perp \in U^\perp$. Notice such decomposition is always unique, so we only need to establish existence.

Lemma 2.1.4

If we can decompose $\vec{v}$ uniquely into $\vec{u}^* + \vec{u}^\perp$ with $\vec{u}^* \in U$ and $\vec{u}^\perp \in U^\perp$, then the decomposition is unique.

Proof. Assume we can find other vectors $\vec{w} \in U$ and $\vec{w}^\perp \in U^\perp$ such that $\vec{v} = \vec{w} + \vec{w}^\perp$, then

$$\vec{u} + \vec{u}^\perp = \vec{w} + \vec{w}^\perp \iff \vec{w}^\perp - \vec{u}^\perp = \vec{u} - \vec{w} \in U \cap U^\perp = \left\{\vec{0}\right\}$$

Hence result. □

In the special case when $B = \{\vec{q}_1, ..., \vec{q}_p, \vec{q}_{p+1}, ..., \vec{q}_n\}$ is an orthonormal basis of $\mathbb{R}^n$ (so that $\vec{q}_i^T \vec{q}_j = 1$ if $i = j$ and 0 otherwise), and the subset $\{\vec{q}_1, ..., \vec{q}_p\}$ is a basis of U, such decomposition always exist. For all $\vec{v} \in \mathbb{R}^n$ there are constants $\beta_1, ..., \beta_n$ such that

$$\vec{v} = \sum_{j=1}^{n} \beta_j \vec{q}_j = \underbrace{\sum_{j=1}^{p} \beta_j \vec{q}_j}_{:=\vec{u}^* \in U} + \underbrace{\sum_{j=p+1}^{n} \beta_j \vec{q}_j}_{:=\vec{u}^\perp} \tag{2.9}$$

Notice that for all $i = 1, ..., p$, we have $\vec{q}_i^T \vec{u}^\perp = 0$, thus we we have $\vec{u}^\perp$ as required. Thus such decomposition exists.

The remaining question is, is the case above too restrictive? Turns out it is not, since there is an algorithm to turn any basis to an orthonormal basis. This is known as the Classical Gram Schmidt (CGS). Let's say we have a vector space spanned by two linear independent vectors $\{\vec{a}_1, \vec{a}_2\}$. Here is the steps of CGS:

1. Normalise $\vec{a}_1$ to obtain $\vec{q}_1 = \vec{a}_1 / \|\vec{a}_1\|$.
2. Decompose $\vec{a}_2$ into a sum of vectors, one parallel to $\vec{q}_1$ and one perpendicular to $\vec{q}_1$. As a reminder, the "perpendicular vector" is:

   $$\vec{v}_2 = \vec{a}_2 - (\vec{a}_2^T \vec{q}_1)\vec{q}_1$$

   which is clearly not equal to zero (otherwise $\vec{a}_1$ is proportional to $\vec{a}_2$.
3. Normalise $\vec{v}_2$ to obtain $\vec{q}_2 := \vec{v}_2 / \|\vec{v}_2\|$.

Then $\vec{q}_1$ is orthonormal to $\vec{q}_2$, and $\{\vec{q}_1, \vec{q}_2\}$ is a basis of Span$(\{\vec{a}_1, \vec{a}_2\})$. So can we generalise to vector spaces spanned by three linear independent vectors $\vec{a}_1$, $\vec{a}_2$, $\vec{a}_3$? The answer is yes! Once we have finished steps (1)-(3) for vectors $\vec{a}_1$, $\vec{a}_2$, to obtain $\vec{q}_1$ and $\vec{q}_2$, we continue by letting

$$\vec{v}_3 = \vec{a}_3 - (\vec{a}_3^T \vec{q}_1)\vec{q}_1 - (\vec{a}_3^T \vec{q}_2)\vec{q}_2$$

We can check that $\vec{v}_3$ is perpendicular to both $\vec{q}_1$ and $\vec{q}_2$, so we can normalise $\vec{v}_3$ to obtain $\vec{q}_3$ and we will have a orthonormal basis for Span$(\vec{a}_1, \vec{a}_2, \vec{a}_3)$.

We may continue by induction to obtain a orthonormal basis for any finite vector space in general. We will have

$$\begin{aligned}
&&\vec{q}_1 &= \vec{a}_1 / \|\vec{a}_1\| \\
\vec{v}_2 &= \vec{a}_2 - (\vec{a}_2^T \vec{q}_1)\vec{q}_1 & \vec{q}_2 &= \vec{v}_2 / \|\vec{v}_2\| \\
\vec{v}_3 &= \vec{a}_3 - (\vec{a}_3^T \vec{q}_1)\vec{q}_1 - (\vec{a}_3^T \vec{q}_2)\vec{q}_2 & \vec{q}_3 &= \vec{v}_3 / \|\vec{v}_3\| \\
&\vdots & &\vdots \\
\vec{v}_k &= \vec{a}_k - \sum_{l=1}^{k-1} (\vec{a}_k^T \vec{q}_l)\vec{q}_l & \vec{q}_k &= \vec{v}_3 / \|\vec{v}_3\| \\
&\vdots & &\vdots
\end{aligned}$$

Therefore we have

Theorem 2.1.5: Classical Gram Schmidt

Given $\{\vec{a_1}, ..., \vec{a_m}\}$ is a basis of certain vector space V. We may construct orthonormal basis $\{\vec{q_1}, ..., \vec{q_m}\}$ such that it spans V, and $\vec{q}_i, \vec{q}_j$ are mutually orthonormal, i.e. $\vec{q}_i^T \vec{q}_j = \delta_{ij}$.

Remark. Worth noting that for all $k \leq m$, $\text{Span}(\vec{a}_1, ..., \vec{a}_k) = \text{Span}(\vec{q}_1, ..., \vec{q}_k)$

which leads to

Corollary 2.1.6: Unique Decomposition of Vectors

Let $U = \text{Span}(\{\vec{u}_1, ..., \vec{u}_p\})$ is a subspace of $\mathbb{R}^n$. Then any $\vec{v} \in \mathbb{R}^n$ can be uniquely decomposed into $\vec{u}^* + \vec{u}^\perp$ with $\vec{u}^* \in U$ and $\vec{u}^\perp \in U^\perp$.

Proof. Extend the set $\{\vec{u}_1, ..., \vec{u}_p\}$ to $\{\vec{u}_1, ..., \vec{u}_n\}$ so that it becomes a basis of $\mathbb{R}^n$, then orthonormalise the new basis using CGS. Then we may use the argument in page 19 to conclude. □

2.1.3 Closed Form of Solution to Least Square Problem

So the map $S := S_U$ sending $\vec{v}$ to $\vec{u}^*$ is well-defined. Moreover, it is linear.

Remark. If $\vec{v} \in U$, then $S_U(\vec{v}) = \vec{v}$. If $\vec{v} \in U^\perp$, then $S_U(\vec{v}) = \vec{0}$.

We will show later that $S_U = T_U$ by showing $S_U(\vec{v})$ minimises the 2-norm of residual vector. But we first obtain closed form of S_U (i.e. find a matrix representing S_U). If we look back to the special case on page 19 when $B = \{\vec{q}_1, ..., \vec{q}_p, \vec{q}_{p+1}, ..., \vec{q}_n\}$ is an orthonormal basis of $\mathbb{R}^n$, and the subset $\{\vec{q}_1, ..., \vec{q}_p\}$ is a basis of U. Then finding $\vec{u^*}$ is equivalent to solving (2.9) for $(\beta_1, ..., \beta_n)$, then extract $\beta_1, ..., \beta_p$ and setting $\vec{u}^* = \sum_{j=1}^p \beta_j \vec{q}_j$. Recall (2.9) is the following

$$\vec{v} = \sum_{j=1}^n \beta_j \vec{q}_j = \underbrace{\sum_{j=1}^p \beta_j \vec{q}_j}_{:=\vec{u}^* \in U} + \underbrace{\sum_{j=p+1}^n \beta_j \vec{q}_j}_{:=\vec{u}^\perp}$$

which can be rewritten as

$$\underbrace{\begin{pmatrix} \uparrow & \uparrow & & \uparrow \\ \vec{q}_1 & \vec{q}_2 & \dots & \vec{q}_n \\ \downarrow & \downarrow & & \downarrow \end{pmatrix}}_{:=Q} \begin{pmatrix} \beta_1 \\ \dots \\ \beta_n \end{pmatrix} = \vec{v} \iff \begin{pmatrix} \beta_1 \\ \dots \\ \beta_n \end{pmatrix} = Q^T \vec{v} \tag{2.10}$$

But we usually don't know $\vec{q}_{p+1}, ..., \vec{q}_n$, and the entries $\beta_{p+1}, ..., \beta_n$ are useless anyway. We may simplify by letting $M_{n\times p} \in \mathbb{R}^{n\times p}$ such that

$$M_{n\times p} = \begin{pmatrix} I^{(p)} \\ O_{(n-p)\times p} \end{pmatrix} = \left(\begin{array}{cccc} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \dots & 1 \\ \hline 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \dots & 0 \end{array}\right) \tag{2.11}$$

also introducing the "thinner" versions of matrix $\hat{Q}$ and its complementary $\hat{Q}'$

$$\hat{Q} = \begin{pmatrix} \uparrow & \uparrow & & \uparrow \\ \vec{q}_1 & \vec{q}_2 & \dots & \vec{q}_p \\ \downarrow & \downarrow & & \downarrow \end{pmatrix}, \quad \hat{Q}^{\perp} = \begin{pmatrix} \uparrow & \uparrow & & \uparrow \\ \vec{q}_{p+1} & \vec{q}_{p+2} & \dots & \vec{q}_n \\ \downarrow & \downarrow & & \downarrow \end{pmatrix} \tag{2.12}$$

so that $Q = (\hat{Q} \,|\, \hat{Q}')$. Then we have the following relations:

$$\hat{Q} = QM_{n\times p} \tag{2.13}$$

$$\begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} = M_{n\times p}^T \begin{pmatrix} \beta_1 \\ \vdots \\ \vdots \\ \beta_n \end{pmatrix} \tag{2.14}$$

So we have

Lemma 2.1.7: Projection for Orthogonal Design

$$S_U(\vec{v}) = \hat{Q}M_{n\times p}^T Q^T \vec{v} = \hat{Q}\hat{Q}^T \vec{v} \tag{2.15}$$

Remark. Using similar argument we can show that $S_{U^{\perp}}(\vec{v}) = (\hat{Q}^{\perp})(\hat{Q}^{\perp})^T \vec{v}$. Moreover $\hat{Q}^T\hat{Q} = I^{(p)}$ and $(\hat{Q}^{\perp})^T\hat{Q}^{\perp} = I^{(n-p)}$.

How can we generalise this case? We aim to express X in terms of Q – this follows directly from CGS:

Lemma 2.1.8: (Thin) QR Factorization

$\exists\, \hat{Q} \in \mathbb{R}^{n\times p}$, $R \in \mathbb{R}^{p\times p}$, R upper-triangular, such that $X = \hat{Q}R$.

Proof. Note by CGS that $\vec{a}_1 = \|\vec{a}_1\|\,\vec{q_1}$ and $\forall 2 \leq k \leq p$,

$$\vec{u}_k = \vec{v}_k + \sum_{l=1}^{k-1}(\vec{u_k}^T\vec{q}_l)\vec{q_l} = \|\vec{v}_k\|\,\vec{q}_k + \sum_{l=1}^{k-1}(\vec{u_k}^T\vec{q}_l)\vec{q_l}$$

From this we know that

$$R = \begin{pmatrix} \|\vec{v}_1\| & \vec{u_2}^T\vec{q}_1 & \dots & \dots & \vec{u_p}^T\vec{q}_1 \\ & \|\vec{v}_2\| & \vec{u_3}^T\vec{q}_2 & \dots & \vec{u_p}^T\vec{q}_2 \\ & & \ddots & & \vdots \\ & & & & \|\vec{v}_p\| \end{pmatrix} \tag{2.16}$$

□

We can therefore find a closed form for $S_U(\vec{v})$:

Theorem 2.1.9

$$S_U(\vec{v}) = P\vec{v}, \qquad P = X(X^TX)^{-1}X^T \tag{2.17}$$

Proof. By noting that $\hat{Q} = XR^{-1} \iff \hat{Q}^T = (R^{-1})^TX^T$, we have

$$\begin{aligned} \hat{Q}\hat{Q}^T = XR^{-1}(R^{-1})^TX^T &= X(R^TR)^{-1}X^T \\ &= X(R^T(\hat{Q}^T\hat{Q})R)^{-1}X^T \\ &= X((\hat{Q}R)^T(\hat{Q}R))^{-1}X^T \\ &= X(X^TX)^{-1}X^T \end{aligned}$$

□

We finally establish that $S_U(\vec{v})$ solves the least square problem, i.e. $\vec{\beta}^* = (X^TX)^{-1}X^T\vec{v}$ minimises $f(\vec{\beta})$. Recall that

$$f(\vec{\beta}) = (\vec{v} - X\vec{\beta})^T(\vec{v} - X\vec{\beta}) = \vec{\beta}^T(X^TX)\vec{\beta} - 2(X^T\vec{v})^T\vec{\beta} + \vec{v}^T\vec{v}$$

which is similar to a quadratic function. Can we "complete the square" to establish that $\vec{\beta}^*$ is the global minimum? Unfortunately it is tricky - if we can establish that

$$f(\vec{\beta}) = \left\|X(\vec{\beta} - \vec{\beta}^*)\right\|^2 + \text{some scalar}$$

then we are done, since we know that a norm of vector is always non-negative, so $f(\vec{\beta}) \geq$ "that scalar" with equality holds iff $X(\vec{\beta} - \vec{\beta}^*) = 0$. But X is full rank so the equality is equivalent to $\vec{\beta} - \vec{\beta}^*$, i.e. the unique minimiser of $f(\vec{\beta})$ is $\vec{\beta}^*$. Our remaining task is unfortunately some tedious algebra.

Notice that

$$\begin{aligned}\left\|X(\vec{\beta}-\vec{\beta}^*)\right\|^2 &= (X(\vec{\beta}-\vec{\beta}^*))^T(X(\vec{\beta}-\vec{\beta}^*))\\ &= (\vec{\beta}^T X^T - \vec{\beta}^{*T}X^T)(X\vec{\beta}-X\vec{\beta}^*))\\ &= \vec{\beta}^T(X^TX)\vec{\beta} - \vec{\beta}^T(X^TX)\vec{\beta}^* - \vec{\beta}^{*T}(X^TX)\vec{\beta} + \vec{\beta}^{*T}(X^TX)\vec{\beta}^*\\ &= \vec{\beta}^T(X^TX)\vec{\beta} - \vec{\beta}^T(X^T\vec{v}) - (X^T\vec{v})^T\vec{\beta} + \vec{v}^TX(X^TX)^{-1}X^T\vec{v}\\ &= \vec{\beta}^T(X^TX)\vec{\beta} - 2(X^T\vec{v})\vec{\beta} + \vec{v}^TX(X^TX)^{-1}X^T\vec{v}\\ &= \vec{\beta}^T(X^TX)\vec{\beta} - 2(X^T\vec{v})\vec{\beta} + \vec{v}^TP\vec{v}\end{aligned}$$

But we know that $P^TP = P^2 = P$. Therefore,

$$f(\vec{\beta}) = \left\|X(\vec{\beta}-\vec{\beta}^*)\right\|^2 + \|(I-P)\vec{v}\|^2 \tag{2.18}$$

which establishes that

Theorem 2.1.10: Solution to Least Square Problem

$\vec{\beta}^*$ is the minimiser of $\vec{\beta}$ - hence the solution to our least square problem is

$$T_U(\vec{v}) = S_U(\vec{v}) = P\vec{v} = X(X^TX)^{-1}X^T\vec{v} \tag{2.19}$$

Exercise 2.1.11: Least Square Problem via Differentiation

You might be tempting to use derivatives to minimise $f(\vec{\beta})$. If so, you might follow the following steps: first establishes

$$\vec{\nabla}f = 2(A\vec{\beta}-\vec{b}) \tag{2.20}$$
$$D^2f = 2A \tag{2.21}$$

where $\vec{\nabla}f$ and D^2f are the gradient (Jacobian) and Hessian of f respectively, $A = X^TX$ and $\vec{b} = X^T\vec{v}$. Convince yourself that A is positive definite. Then we have shown that f is strictly convex must have a unique minimiser. The minimiser $\vec{\beta}^*$ is a critical point so must satisfy $\vec{\nabla}f = \vec{0}$ which has one solution.

2.1.4 Digression I: Properties of Projection Matrix

We take a step back to look at some properties of matrix P. Firstly, we establish well-definedness of P by ensuring that X^TX is invertible.

Exercise 2.1.12

Assume $X \in \mathbb{R}^{n\times p}$ with $n > p$ with rank $r \leq p$. Show that

1. $\ker(X^TX) = \ker(X)$
2. $\text{rank}(X^TX) = \text{rank}(X)$
3. (X^TX) invertible $\iff$ $\text{rank}(X) = p$

Hint. For (1), think about $\vec{c}^T X^T X \vec{c} = \|X\vec{c}\|^2$. For (2), think about rank-nullity.

Secondly, notice that $P^2 = P$ (idempotent) and $P^T = P$ (symmetric).

Exercise 2.1.13: Idempotence and Symmetry of P

Using (2.17), show that $P^2 = P$ and $P^T = P$.

We called any matrices satisfying the above properties *projection matrix*. Notice that by spectral theorem, the eigenvalues of P are real, and P exhibits a spectral decomposition $P = \Gamma D \Gamma^T$ with Γ orthogonal. In fact using similar argument in Exercise 2.1.1 that the only eigenvalues of P are 0 and 1.

What is the spectral decomposition of P, then? It is hidden in the formula (2.15):

$$P = \hat{Q}\hat{Q}^T = QM_{n\times p}M_{p\times n}^T Q^T = \underbrace{Q}_{\Gamma} \underbrace{\begin{pmatrix} I^{(p)} & \\ & O_{(n-p)\times(n-p)} \end{pmatrix}}_{D} \underbrace{Q^T}_{\Gamma^T} \tag{2.22}$$

So the (algebraic) multiplicities of eigenvalues of 0 and 1 are p and $n-p$ respectively. Moreover, we have $\text{tr}(P) = \text{rank}(P) = p$.

Remark. One might wonder whether any projection matrices P defines a projection. The answer is yes - if $\text{Csp}(P)$ represents the column span of P, then we have $T_{\text{Csp}(P)}(\vec{v}) = P\vec{v}$. Just check that

- $\forall \vec{x} \in \text{Csp}(P), \exists \vec{z} \in \mathbb{R}^n$ such that $\vec{x} = P\vec{z}$ ($\vec{x}$ is a linear combination of columns of P). Hence $P\vec{x} = P^2\vec{z} = P\vec{z} = \vec{x}$.
- Given $\vec{x} \in \text{Csp}(P)^\perp$, then $\forall \vec{y} \in \mathbb{R}^n$, $(P\vec{x})^T\vec{y} = \vec{x}^T P^T \vec{y} = \vec{x}^T P\vec{y} = 0$. Hence $P\vec{x} = \vec{0}$.

Recall that P admits a spectral decomposition $\Gamma D \Gamma^T$. It should be easy to show that if $r = \text{rank}(P)$, then the number of 0 and 1 in D are $n - r$ and r respectively, since the number of 1 in D equals to the rank of P. Therefore we have $\text{tr}(P) = r$.

A final remark is that we can show that

$$(\hat{Q}^\perp)(\hat{Q}^\perp)^T = Q \begin{pmatrix} O_{p\times p} & \\ & I^{(n-p)} \end{pmatrix} Q^T = I - P \tag{2.23}$$

So we have

$$T_{U^\perp}(\vec{v}) = (I - P)\vec{v} \tag{2.24}$$

Example 2.1.14: A Geometrical Problem

Let $U \leq \mathbb{R}^3$ be the plane spanned by the vectors $(-2, -3, 1)$ and $(1, -2, -2)$. We want to find the perpendicular distance from the point $(-1, -3, -2)$ to U.

Usual way: First find the normal of plane by cross product

$$(-2, -3, -1) \times (1, -2, -2) = (4, -5, 7)$$

The perpendicular distance is then the length of orthogonal projection of vector $(-1,-3,-2)$ to U, which is $1/\sqrt{10}$.

via Theorem 2.1.10: Let $X = \begin{pmatrix} -2 & 1 \\ -3 & -2 \\ -1 & -2 \end{pmatrix}$ Then

$$X^TX = \begin{pmatrix} 14 & 6 \\ 6 & 9 \end{pmatrix} \Rightarrow (X^TX)^{-1} = \frac{1}{90}\begin{pmatrix} 9 & -6 \\ -6 & 14 \end{pmatrix}$$

Therefore

$$\begin{aligned} P = X(X^TX)^{-1}X^T &= \frac{1}{90}\begin{pmatrix} -2 & 1 \\ -3 & -2 \\ -1 & -2 \end{pmatrix}\begin{pmatrix} 9 & -6 \\ -6 & 14 \end{pmatrix}\begin{pmatrix} -2 & -3 & -1 \\ 1 & -2 & -2 \end{pmatrix} \\ &= \frac{1}{90}\begin{pmatrix} 74 & 20 & -28 \\ 20 & 65 & 35 \\ -28 & 35 & 41 \end{pmatrix} \end{aligned}$$

Thus the orthogonal projection of $(-1,-3,-2)$ onto the orthogonal set of U, which is

$$\begin{aligned} \vec{u} = (I-P)\begin{pmatrix} -1 \\ -3 \\ -2 \end{pmatrix} &= \frac{1}{90}\begin{pmatrix} 16 & -20 & 28 \\ -20 & 25 & -35 \\ 28 & -35 & 49 \end{pmatrix}\begin{pmatrix} -1 \\ -3 \\ -2 \end{pmatrix} \\ &= \frac{1}{90}\begin{pmatrix} -12 \\ 15 \\ -21 \end{pmatrix} = \frac{1}{30}\begin{pmatrix} -4 \\ 5 \\ -7 \end{pmatrix} \end{aligned}$$

The shortest distance required is thus $\|\vec{u}\| = \frac{1}{30}\sqrt{(-4)^2+5^2+(-7)^2} = \frac{1}{\sqrt{10}}$ as required. (Observe the basis Csp(P) is in fact the multiple of normal as obtained by cross products.)

2.2 Ordinary Least Square Estimator (OLSE)

Now we can define a least square estimator for any linear model.

Definition 2.2.1: Ordinary Least Square Estimator (OLSE)

If we have a linear model of $(\vec{X}_1, \vec{X}_2, ..., \vec{X}_n, \vec{Y})$ with $\vec{Y} = (Y_1, ..., Y_n)$ satisfies

$$\vec{Y} \mid \vec{X}_1, ..., \vec{X}_n \sim X\vec{\beta} + \vec{\epsilon} \tag{2.25}$$

where X is defined in (1.5), then we define the least square estimator as a function $\hat{\beta}_{\vec{X}_1,...,\vec{X}_n} : \mathbb{R}^n \to \mathbb{R}^p$ by

$$\hat{\beta} := \hat{\beta}_{\vec{X}_1,...,\vec{X}_n}(\vec{y}) = (X^TX)^{-1}X^T\vec{y}, \quad \vec{y} = (y_1, ..., y_n) \tag{2.26}$$

Remark. Technically $\hat{\beta}$ is a function from $\mathbb{R}^{n(d+1)}$ to $\mathbb{R}^p$, but we have assumed from our heuristics that $\vec{X}_1, ..., \vec{X}_n$ are "fixed". Also note that $\hat{\beta}$ is linear in $\vec{y}$.

Example 2.2.2: Formula for OLSE of Simple (Normal) Linear Model

Recall that a simple linear model of data $\left\{(\vec{X}_i, Y_i)\right\}_{i=1}^n$ can be written as

$$Y_i \,|\, X_i \sim \beta_1 + \beta_2 X_i + \epsilon_i, \quad \epsilon_i$$

(if ϵ_i are iid $\mathbf{N}(0, \sigma^2)$ then the model is linear), which can be rewritten as

$$\vec{Y} \,|\, \vec{X} \sim \underbrace{\begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix}}_{:=X} \underbrace{\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}}_{:=\vec{\beta}} + \vec{\epsilon}, \qquad \vec{\epsilon} \sim \mathbf{N}_n(0, \sigma^2 I^{(1078)})$$

The least square estimates of $\vec{\beta}$ is $\hat{\beta}(\vec{Y}) = (X^T X)^{-1} X^T \vec{Y}$ (sorry for the abuse of notation). Note that

$$X^T X = \begin{pmatrix} 1 & 1 & ... & 1 \\ X_1 & X_2 & ... & X_n \end{pmatrix} \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ ... & ... \\ 1 & X_n \end{pmatrix} = \begin{pmatrix} n & \sum_{i=1}^n X_i \\ \sum_{i=1}^n X_i & \sum_{i=1}^n X_i^2 \end{pmatrix}$$

$$\Rightarrow (X^T X)^{-1} = \frac{1}{n\sum_{i=1}^n X_i^2 - \left(\sum_{i=1}^n X_i\right)^2} \begin{pmatrix} \sum_{i=1}^n X_i^2 & -\sum_{i=1}^n X_i \\ -\sum_{i=1}^n X_i & n \end{pmatrix}$$

Moreover we have $X^T \vec{Y} = \begin{pmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n X_i Y_i \end{pmatrix}$ Therefore

$$\hat{\beta} = \begin{pmatrix} \frac{\sum_{i=1}^n X_i^2 \sum_{i=1}^n Y_i - \sum_{i=1}^n X_i \sum_{i=1}^n X_i Y_i}{n\sum_{i=1}^n X_i^2 - \left(\sum_{i=1}^n X_i\right)^2} \\ \frac{n\sum_{i=1}^n X_i Y_i - \sum_{i=1}^n X_i \sum_{i=1}^n Y_i}{n\sum_{i=1}^n X_i^2 - \left(\sum_{i=1}^n X_i\right)^2} \end{pmatrix} \tag{2.27}$$

Let

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}, \quad \bar{Y} = \frac{\sum_{i=1}^n Y_i}{n}$$

$$S_{xx} = \sum_{i=1}^n (X_i - \bar{X})^2 = n s_x^2, \quad S_{xy} = \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$

then

$$\hat{\beta}_2 = \frac{n\sum_{i=1}^n X_i Y_i - \sum_{i=1}^n X_i \sum_{i=1}^n Y_i}{n\sum_{i=1}^n X_i^2 - \left(\sum_{i=1}^n X_i\right)^2} = \frac{\sum_{i=1}^n X_i Y_i - \frac{\sum_{i=1}^n X_i \sum_{i=1}^n Y_i}{n}}{\sum_{i=1}^n X_i^2 - \frac{(\sum_{i=1}^n X_i)^2}{n}} = \frac{S_{xy}}{S_{xx}}$$

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^n X_i^2 \sum_{i=1}^n Y_i - \sum_{i=1}^n X_i \sum_{i=1}^n X_i Y_i}{n\sum_{i=1}^n X_i^2 - \left(\sum_{i=1}^n X_i\right)^2} \\
&= \frac{\bar{Y}\sum_{i=1}^n X_i^2 - \bar{X}\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2 - \frac{(\sum_{i=1}^n X_i)^2}{n}} \\
&= \frac{\bar{Y}\sum_{i=1}^n X_i^2 - \bar{Y}\frac{(\sum_{i=1}^n X_i)^2}{n} - \bar{X}\sum_{i=1}^n X_i Y_i + \bar{Y}\frac{(\sum_{i=1}^n X_i)^2}{n}}{\sum_{i=1}^n X_i^2 - \frac{(\sum_{i=1}^n X_i)^2}{n}} \\
&= \bar{Y} - \bar{X}\frac{\sum_{i=1}^n X_i Y_i - \frac{(\sum_{i=1}^n X_i)(\sum_{i=1}^n Y_i)}{n}}{\sum_{i=1}^n X_i^2 - \frac{(\sum_{i=1}^n X_i)^2}{n}} \\
&= \bar{Y} - \hat{\beta}_1 \bar{X}
\end{aligned}$$

to summarise

$$\hat{\beta}_2 = \frac{S_{xy}}{S_{xx}} \tag{2.28}$$

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X} \tag{2.29}$$

Remark. It is interesting to see how (2.28) and (2.29) resembles (1.4). The reason behind will be discussed in the second part of the book.

Example 2.2.3: Data Example I (Part II) - Evaluating OLSE

Followed from the previous example we have the following

```
Y = data.Son
X = hcat(ones(length(data.Father)), data.Father)
```

We may use (2.26) to obtain our OLSE.

```
betahat = (X'*X) \ (X'*Y)
```

```
2-element ArrayFloat64,1:
 33.89280054067077
  0.514005912545458
```

so we have $\hat{\beta}_1 \approx 33.893$ and $\hat{\beta}_2 \approx 0.514$. Here is a plot of fitted line on the original scatter plot.

```
using Plots # or StatsPlots, but Plots is sufficient here
scatter(data.Father,data.Son,...)
plot!(x -> betahat[1] + betahat[2]*x, ...)
```

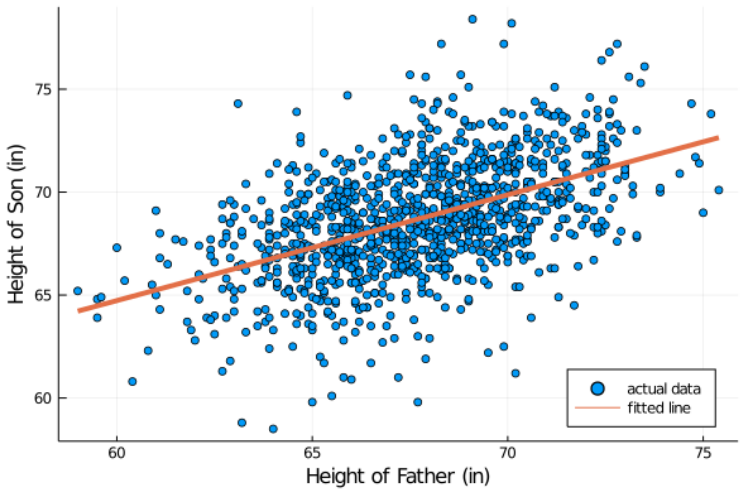

Figure 2.2: Fitted Line with Scatter Plot

2.2.1 Quality of OLSE

Two ways to assess the properties of OLSE are calculating its bias and covariance. Ideally we want to reduce both of them. Recall that

$$\text{bias}(\hat{\beta}) = \mathbb{E}(\hat{\beta}(\vec{Y}) \mid X) - \vec{\beta} \tag{2.30}$$

Proposition 2.2.4: Bias and Variance of OLSE

$\text{bias}(\hat{\beta}) = \vec{0}$ and $\text{Cov}(\hat{\beta}(\vec{Y}) \mid X) = \sigma^2(X^TX)^{-1}$

Proof. Notice that $\vec{Y}$ has mean $\vec{X}\vec{\beta}$ and covariance $\sigma^2 I^{(n)}$. Therefore $\mathbb{E}(\hat{\beta}(\vec{Y})|X) = (X^TX)^{-1}X^T\mathbb{E}(\vec{Y}) = (X^TX)^{-1}X^T(X\vec{\beta}) = \vec{\beta}$ and $\text{bias}(\hat{\beta}) = \vec{0}$. Moreover,

$$\begin{aligned}
\text{Cov}(\hat{\beta}|X) &= (X^TX)^{-1}X^T\text{Cov}(\vec{Y}|\vec{X})((X^TX)^{-1}X^T)^T \\
&= (X^TX)^{-1}X^T(\sigma^2 I_n)X((X^TX)^{-1})^T \\
&= \sigma^2(X^TX)^{-1}(X^TX)^T((X^TX)^{-1})^T \\
&= \sigma^2(X^TX)^{-1}((X^TX)^{-1}(X^TX))^T = \sigma^2(X^TX)^{-1}
\end{aligned}$$

Hence result. □

In practice we need to estimate σ^2 as well:

Proposition 2.2.5: Unbiased Estimator of σ^2

Write the Sum of Squared Errors (SSE) as

$$\text{SSE} := \text{SSE}(X, \vec{y}) = \left\|\vec{y} - X\hat{\beta}(\vec{y})\right\|^2 = \left\|(I^{(n)} - P)\vec{y}\right\|^2 \tag{2.31}$$

Then $\hat{\sigma}^2$ is an unbiased estimator of σ^2, where

$$\hat{\sigma}^2 = \frac{\text{SSE}}{n-p} \tag{2.32}$$

Proof. It is important to note that

$$\text{SSE}(X, \vec{y}) = \text{tr}((I^{(n)} - P)\vec{y}\,((I^{(n)} - P)\vec{y})^T) \tag{2.33}$$

and therefore

$$\begin{aligned}
\mathbb{E}(\text{SSE}(X, \vec{Y}) \mid X) &= \mathbb{E}(\text{tr}((I^{(n)} - P)\vec{Y}\,((I^{(n)} - P)\vec{Y})^T) \mid X) \\
&= \text{tr}(\text{Cov}((I^{(n)} - P)\vec{Y} \mid X)) \\
&= \text{tr}((I^{(n)} - P)\sigma^2 I^{(n)}(I^{(n)} - P)^T) \\
&= \sigma^2\,\text{tr}(I^{(n)} - P) = \sigma^2(n-p)
\end{aligned}$$

□

Remark. Consider a model with intercept, i.e. $\vec{Y}|\vec{X} \sim X\vec{\beta} + \vec{\epsilon}$ with X **contain a column of all ones** $\mathbf{1}_n$ (without loss of generality assume the first column of X is $\mathbf{1}_n$). Then the constant model $\vec{Y}|\vec{X} \sim \mathbf{1}_n\mu + \vec{\epsilon}$ can be viewed as a simplification (or "sub-model") of original model.

We write the Sum of Squared Total (SST) as

$$\text{SST} := \text{SST}(\vec{y}) = \text{SSE}(\mathbf{1}_n, \vec{y}) = \|\vec{y} - \mathbf{1}_n\bar{y}\|^2 \tag{2.34}$$

where $\bar{y}$ is the sample mean of entries of $\vec{y}$ (notice it is the OLSE of the constant model with $X = \mathbf{1}_n$, the vector with all entries 1.) Also define the Sum of Squares Regression (SSR) as

$$\text{SSR} := \text{SSR}(X, \vec{y}) = \left\|X\hat{\beta}(\vec{y}) - \mathbf{1}_n\bar{y}\right\|^2 \tag{2.35}$$

For models with intercept, the following decomposition holds

$$\text{SST} = \text{SSE} + \text{SSR} \tag{2.36}$$

and therefore SST is always greater than SSE. In particular, if $\vec{Y} = X\vec{\beta} + \vec{\epsilon}$ then we expect SSE $\ll$ SST. Indeed, if we define

$$R^2 = \frac{\text{SST} - \text{SSE}}{\text{SST}} = \frac{\text{SSR}}{\text{SST}} \tag{2.37}$$

Then we expect $R^2 \approx 1$ if $\vec{Y} = X\vec{\beta} + \vec{\epsilon}$ is a good fit. We will devise statistical test for this in the next chapter.

Exercise 2.2.6

Prove (2.36) for the case when the model has intercept (X has a column $\mathbf{1}_n$).

Hint. Notice that $\mathbf{1}_n \in \mathrm{CSp}(X) = \mathrm{CSp}(P)$ so $P\mathbf{1}_n = \mathbf{1}_n$. Therefore,

$$\begin{aligned}\mathrm{SSE} + \mathrm{SSR} &= \left\|(I^{(n)} - P)\vec{y}\right\|^2 + \|P\vec{y} - \mathbf{1}_n\bar{y}\|^2 \\ &= \vec{y}^T(I^{(n)} - P)^T(I^{(n)} - P)\vec{y} + \vec{y}^T P^T P\vec{y} - 2\bar{y}\,\mathbf{1}_n^T P\vec{y} + (\mathbf{1}_n^T\mathbf{1}_n)\,\|\bar{y}\|^2 \\ &= \vec{y}^T(I^{(n)} - P)\vec{y} + \vec{y}^T P\vec{y} - 2\bar{y}(P\mathbf{1}_n)^T\vec{y} + n\,\|\bar{y}\|^2 \\ &= \vec{y}^T\vec{y} - 2\bar{y}\mathbf{1}_n^T\vec{y} + n\,\|\bar{y}\|^2 = \|\vec{y} - \bar{y}\mathbf{1}_n\|^2 = \mathrm{SST}\end{aligned}$$

The significance of OLSE is that it is the minimiser of covariance among unbiased linear estimator (linear w.r.t. $\vec{y}$). In fact we can prove a stronger result.

Theorem 2.2.7: Gauss-Markov Theorem

Let $\vec{c} \in \mathbb{R}^p$. Then $\vec{c}^T\hat{\beta}(\vec{y})$ is an unbiased, linear estimator of $\vec{c}^T\vec{\beta}$. Moreover, it is the minimiser of covariance among unbiased linear estimator of $\vec{c}^T\vec{\beta}$.

Proof. Suppose $\hat{\gamma}(\vec{y}) = \vec{L}^T\vec{y}$ is another unbiased linear estimator of $\vec{c}^T\vec{\beta}$. We would like to find $\mathrm{Cov}(\hat{\gamma}(\vec{Y})) = \mathrm{Cov}(\vec{L}^T\vec{Y})$ [a]. We begin by noting

$$\mathrm{Cov}(\vec{L}^T\vec{Y}) = \mathrm{Cov}(\vec{c}^T\hat{\beta} + \vec{L}^T\vec{Y} - \vec{c}^T\hat{\beta}) = \mathrm{Cov}(\vec{c}^T\hat{\beta} + (\vec{L}^T - \vec{c}^T(X^TX)^{-1}X^T)\vec{Y})$$

Let $\vec{D} := \vec{L} - X((X^TX)^{-1})^T\vec{c}$. Then

$$\mathrm{Cov}(\hat{\gamma}) = \mathrm{Cov}(\vec{c}^T\hat{\beta} + \vec{D}^T\vec{Y}) = \mathrm{Cov}(\vec{c}^T\vec{\beta}) + \mathrm{Cov}(\vec{D}^T\vec{Y}) + 2\mathrm{Cov}(\vec{c}^T\vec{\beta}, \vec{D}^T\vec{Y})$$

Observe that $\mathbb{E}(\hat{\gamma}) = \mathbb{E}(\vec{c}^T\hat{\beta}) = \vec{c}^T\vec{\beta}$ (since they are unbiased). Therefore $\mathbb{E}(\vec{D}^T\vec{Y}) = \mathbb{E}(\vec{L}^T\vec{Y} - \vec{c}^T\hat{\beta}) = \mathbb{E}(\hat{\gamma}) - \mathbb{E}(\vec{c}^T\hat{\beta}) = \vec{0}$. But $\mathbb{E}(\vec{D}^T\vec{Y}) = \vec{D}^T\mathbb{E}(\vec{Y}) = \vec{D}^TX\vec{\beta}$. Therefore $\vec{D}^TX = \vec{0}^T$, or $X^T\vec{D} = (\vec{D}^TX)^T = \vec{0}$. Finally,

$$\begin{aligned}\mathrm{Cov}(\vec{c}^T\vec{\beta}, \vec{D}^T\vec{Y}) = \mathrm{Cov}(\vec{c}^T(X^TX)^{-1}X^T\vec{Y}, \vec{D}^T\vec{Y}) &= \vec{c}^T(X^TX)^{-1}X^T\mathrm{Cov}(\vec{Y}, \vec{Y})\vec{D} \\ &= \vec{c}^T(X^TX)^{-1}X^T(\sigma^2 I_n)\vec{D} \\ &= \vec{c}^T(X^TX)^{-1}X^T\vec{D}\sigma^2 = \vec{0}\end{aligned}$$

Therefore $\mathrm{Cov}(\hat{\gamma}) = \mathrm{Cov}(\vec{c}^T\hat{\beta}) + \mathrm{Cov}(\vec{L}^T\vec{Y}) \geq \mathrm{Cov}(\vec{c}^T\hat{\beta}) = \vec{c}^T(X^TX)^{-1}\vec{c}\sigma^2$ □

[a] I am too lazy to include the conditional sign now...

Why we need to think about linear combinations of entries of $\vec{\beta}$? There are a few scenarios we may come across:

- You want to estimate a coefficient, say $\beta_1 = (1, 0., ..., 0)^T\vec{\beta}$.
- You want to estimate the difference of two coefficients, say $\beta_1 - \beta_2 = (1, -1., ..., 0)^T\vec{\beta}$.
- You want to make prediction, i.e. given covariates $\vec{x} = x_1, ..., x_d$, you

want to predict the expectation $\mathbb{E}(Y \mid \vec{X} = \vec{x}) = f_1(\vec{x})\beta_1 + ... + \beta_p f_p(\vec{x}) = (f_1(\vec{x}), ..., f_p(\vec{x}))^T\vec{\beta}$.

All three cases are covered by the Gauss-Markov Theorem, saying that the estimators $\vec{c}^T\hat{\beta}$ (with an appropriate choice of $\vec{c}$) are the unbiased linear estimator minimising its covariance. We say that $\vec{c}^T\hat{\beta}$ is the *Best Linear Unbiased Estimator (BLUE)* of $\vec{c}^T\vec{\beta}$.

Example 2.2.8: Data Example I (Part III) - Interpretation of Gauss Markov Theorem

By Gauss Markov Theorem:

- $\hat{\beta}_1$ (as defined in (2.29)) is the BLUE of β_1
- $\hat{\beta}_2$ (as defined in (2.28)) is the BLUE of β_2
- $\hat{\beta}_1 + 65\hat{\beta}_2$ is the BLUE of expected height of son when his father is 65 inches tall.

Remark. What is the variance of the prediction of the *actual* height of son? Notice the prediction is

$$\hat{Y} \mid (\text{height} = 65) = \hat{\beta}_1 + 65\hat{\beta}_2 + \epsilon, \quad \epsilon \sim \mathbf{N}(0, \sigma^2) \tag{2.38}$$

(with an extra ϵ variable to account the deviation of the Son's height from the expectation!) So the variance of prediction of actual height is increased by σ^2 in our case. In general, the variance of *actual* prediction is $\sigma^2(\vec{c}^T(X^TX)^{-1}\vec{c}+1)$, while the variance of *expected* prediction is $\sigma^2(\vec{c}^T(X^TX)^{-1}\vec{c})$.

CHAPTER 3

Inference of Normal Linear Model

So far we haven't assumed the distribution of $\vec{\epsilon}$ in a linear model $\vec{Y} \mid \vec{X} \sim X\vec{\beta} + \vec{\epsilon}$, so let us now assume NTA (i.e. $\vec{\epsilon} \sim \mathbf{N}_n(0, \sigma^2 I^{(n)})$).

Remark. The likelihood of data $\vec{Y}$ is

$$L(\vec{\beta}, \sigma^2; \vec{y}) = f_{\vec{Y} \mid \vec{X}}(\vec{y} \mid \vec{x}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}(\vec{y} - X\vec{\beta})^T(y - \vec{X}\vec{\beta})\right) \quad (3.1)$$

The maximum likelihood estimator $(\hat{\beta}_{\text{MLE}}, \hat{\sigma}^2_{\text{MLE}})$ maximises the likelihood, $L(\vec{\beta}, \sigma^2; \vec{y})$, or the log-likelihood

$$\ell(\vec{\beta}, \sigma^2; \vec{y}) = \ln L(\vec{\beta}, \sigma^2; \vec{y}) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\vec{y} - X\vec{\beta})^T(\vec{y} - X\vec{\beta}) \quad (3.2)$$

But we know from Chapter 2 that for all $\sigma^2 \geq 0$,

$$\ell(\vec{\beta}, \sigma^2; \vec{y}) \leq \ell(\hat{\beta}, \sigma^2; \vec{y}) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{\text{SSE}}{2\sigma^2}$$

where $\hat{\beta}$ is the OLSE. So it is sufficient to maximise $\ell(\sigma^2) := \ell(\hat{\beta}, \sigma^2; \vec{y})$. Notice that

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{\text{SSE}}{2(\sigma^2)^2}$$

So by looking at sign change of $\frac{\partial \ell}{\partial \sigma^2}$ we know that $\ell(\sigma^2)$ reaches maximum at $\sigma^2 = \hat{\sigma}^2_{\text{MLE}} = \text{SSE}/n$. The main point is that the OLSE $\hat{\beta}$ actually maximises the likelihood. When $\vec{\epsilon}$ follows other distributions then we generally study the MLE of $\vec{\beta}$ as we have done here.

Now we have specified the distribution of $\vec{\epsilon}$ (hence $\vec{Y} \mid \vec{X}$) and the estimator $\hat{\beta}(X, \vec{y})$. We can now do what (theoretical) statisticians would like to do – to study the distribution $\hat{\beta}(X, \vec{Y}) \mid \vec{X}$ (the distribution when we plug in the random variable $\vec{Y}$ into the estimator itself), and construct hypothesis tests etc.

3.1 Numerical Experiment

A big theorem in statistics is often motivated by (Monte-Carlo) experiments. In our case, we can get a rough idea of the distribution $\hat{\beta}(X, \vec{Y}) \,|\, \vec{X}$ by generating large samples of $(\vec{X}, \vec{Y})$, plug into the estimator $\hat{\beta}$ and plot density / histogram of final result. Here we outline the procedure in Julia. Let us assume the following model:

$$Y_i \,|\, X_i = 2 + X_i + \epsilon_i, \quad \epsilon_i \sim \mathbf{N}(0, 1)$$

so that

$$\vec{Y} \,|\, \vec{X} = \begin{pmatrix} \uparrow & \uparrow \\ \mathbf{1}_n & \vec{X} \\ \downarrow & \downarrow \end{pmatrix} \quad \vec{\epsilon} \sim \mathbf{N}(0, I^{(n)})$$

We study the distribution of $\vec{Y}$ for varying n.

Step 1: Generate a fixed $\vec{x}$. Here we want $\vec{x}$ to be Gaussian (it can be uniform or whatever since it won't affect the model itself). For better resolution we want our simulated $\vec{x}$ has mean 0 and standard deviation 5, so that most of the points in $\vec{x}$ falls between -10 and 10.

```
x = 5*randn(n)
```

Step 2: Now we can simulate $\vec{y}$ from $\vec{Y} \,|\, \vec{X} = \vec{x}$,

```
y = 2 .+ x .+ randn(n)
```

construct the design matrix, and

```
X = hcat(ones(length(x)), x)
```

plug into our OLSE formula.

```
betahat = (X' * X) \ (X' * Y)
```

We repeat *Step 2* for many times, say 1000.

```
using StatsPlots # using Random
function OLSE_monte_carlo(n; itr=1000)
    # n -- sample size, length of y
    # itr -- number of samples you want to take
    # Random.seed(1234) -- for reproducibility, need Random
    x = 5*rand(n)
    betahat1_arr = ArrayFloat64(undef,itr)
    betahat2_arr = ArrayFloat64(undef,itr)
    for i=1:itr
        Y = 2 .+ x .+ randn(n)
        X = hcat(ones(length(x)), x)
```

```
        betahat1, betahat2 = (X' * X)  (X' * Y)
        betahat1_arr[i] = betahat1
        betahat2_arr[i] = betahat2
    end
    return betahat1_arr, betahat2_arr
end
```

The following contains plot of individual entries of $\hat{\beta}$ for $n = 5, 50$ and 500.

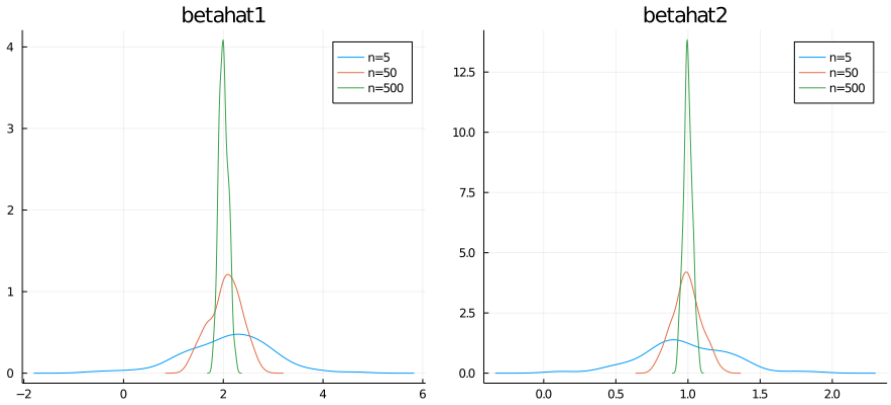

Figure 3.1: Empirical Density Plot when $n = 5, 50$ and 500

The distribution seems to be symmetric, so one may guess they are normally distributed. We may plot a QQ-plot to check:

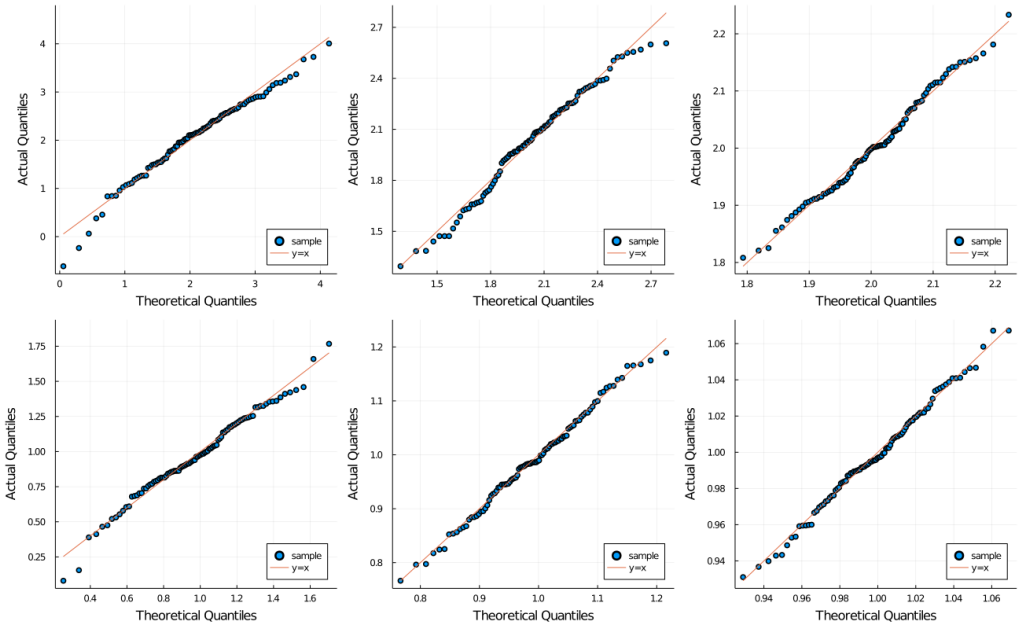

Figure 3.2: Upper Row: QQ-plot for first (upper row) and second (lower row) entry of $\hat{\beta}$, $n = 5, 50, 500$.

When n is small the empirical distribution seems not to fit normal distribution too well (it is heavy tailed). We might guess it follows a t-distribution instead.

3.2 Fischer-Cochran Theorem and t-test

Let us now proceed to the theoretical result. We prove the Fischer-Cochran Theorem and show that linear combination of entries of $\hat{\beta}$ (and, of course, the entries itself) follow a Student-t distribution. We start by a lemma in linear algebra.

Lemma 3.2.1

Let $A_1, ..., A_k$ be symmetric $n \times n$ matrices such that $\sum_{i=1}^k A_i = I^{(n)}$ and rank$A_i = r_i$. Then the following are equivalent:

1. $\sum_{i=1}^k r_i = n$
2. $A_i A_j = 0$ whenever $i \neq j$
3. A_i is idempotent (hence a projection matrix) for all $i = 1, ..., k$

Proof. (2) $\implies$ (3): Just consider $A_i^2 = A_i(\sum_{j=1}^k A_j) = A_i I^{(n)} = A_i$.
(3) $\implies$ (1): Since A_i are projection matrices, by spectral decomposition we have rank$A_i =$ trA_i. Therefore

$$n = \text{tr}(I^{(n)}) = \text{tr}\left(\sum_{i=1}^k A_k\right) = \sum_{i=1}^k \text{tr}(A_k) = \sum_{i=1}^k \text{rank}(A_k) = \sum_{i=1}^k r_i$$

(1) $\implies$ (2): Let B_i be the basis of $V_i = \text{CSp}(A_i)$, and let $B = \cup B_i$. By (1), B has at most n elements. Moreover, notice $\forall \vec{x} \in \mathbb{R}^n$

$$\vec{x} = I^{(n)}\vec{x} = \sum_{i=1}^k A_k \vec{x}$$

So indeed B span $\mathbb{R}^n$. Therefore B (all columns from matrices A_i) is a basis of n. Indeed, we have $\mathbb{R}^n = \bigoplus_{i=1}^k V_i$, the direct sum of spaces V_i. This means $\forall \vec{x} \in \mathbb{R}^n$, there exists uniquely $\vec{v}_i \in V_i$ such that $\vec{x} = \sum_{i=1}^k \vec{v}_i$. It follows that by uniqueness, for all columns of A_j, say $\vec{y} \in V_j$, we have

$$\vec{v} = \vec{0} + ... + \vec{0} + \vec{y} + \vec{0} + ... + \vec{0} = A_1\vec{v} + ... + A_{j-1}\vec{v} + A_j\vec{v} + A_{j+1}\vec{v} + ... + A_n\vec{v}$$

So whenever $i \neq j$ we have $A_j\vec{v} = \vec{0}$ by uniqueness of decomposition. □

We are in a good position to prove the most important theorem in the chapter.

Theorem 3.2.2: Fischer-Cochran Theorem

Let $A_1, ..., A_k$ be $n \times n$ projection matrices with rank$A_i = r_i$ and $\sum_{i=1}^n A_i = I^{(n)}$. If $\vec{Z} \sim \mathbf{N}_n(\vec{\mu}, I^{(n)})$, then $\vec{Z}^T A_1 \vec{Z}, ..., \vec{Z}^T A_k \vec{Z}$ are independent and

$$\vec{Z}^T A_i \vec{Z} \sim \chi^2_{r_i}(\delta_i), \quad \delta_i = \vec{\mu}^T A_i \vec{\mu} \tag{3.3}$$

Proof. We first prove independence. Let $i \neq j$. Then we have $A_i\vec{Z}$ and $A_j\vec{Z}$ jointly normally distributed and $\text{Cov}(A_i\vec{Z}, A_j\vec{Z}) = A_i\text{Cov}(\vec{Z}, \vec{Z})A_j^T = A_i A_j = 0$

(i.e. uncorrelated) by previous lemma. Thus $A_i\vec{Z}$ and $A_j\vec{Z}$ are independent, and so is $\vec{Z}^T A_i \vec{Z}$ and $\vec{Z}^T A_j \vec{Z}$. This proves independence.

It remains to establish the distribution of individual $\vec{Z}^T A_i \vec{Z}$. Recall that that A_i has the spectral decomposition

$$A_i = Q_i \begin{pmatrix} I^{(r_i)} & \\ & O_{(n-r_i)\times(n-p)} \end{pmatrix} Q_i^T$$

with Q orthogonal. If we write $Q_i = (\hat{Q}_i \,|\, \hat{Q}_i^\perp)$, where $\hat{Q}_i$ is the first r_i columns of Q, then we have $A_i = \hat{Q}_i \hat{Q}_i^T$ and $\hat{Q}_i^T \hat{Q} = I^{(p)}$. It follows that

$$\vec{Z}^T A_i \vec{Z} = \left\| \hat{Q}_i \vec{Z} \right\|^T \sim \chi^2_n \left(\left\| \hat{Q}_i \vec{\mu} \right\|^2 \right) = \chi^2_{r_i}(\vec{\mu}^T A_i \vec{\mu}) \tag{3.4}$$

□

We immediate have the following consequence

Theorem 3.2.3: χ^2 test and t test.

Let $\vec{Y}|\vec{X} = X\vec{\beta} + \vec{\epsilon}$ with $\vec{\epsilon} \sim \mathbf{N}_n(0, \sigma^2 I^{(n)})$ and X full rank. Define $\hat{\beta} = (X^T X)^{-1} X^T \vec{Y}$ as the OLSE, and let $\vec{c} \in \mathbb{R}^p$ be a non-zero vector. Then:

1. $\vec{c}^T \hat{\beta}(\vec{Y})$ and $\mathrm{SSE}(X, \vec{Y})$ are independent.
2. $\vec{c}^T \hat{\beta} \sim \mathbf{N}(\vec{c}^T \vec{\beta}, \vec{c}^T (X^T X)^{-1} \vec{c} \sigma^2)$.
3. $\dfrac{\mathrm{SSE}(X, \vec{Y})}{\sigma^2} \sim \chi^2_{n-p}$
4. $\dfrac{\vec{c}^T \hat{\beta} - \vec{c}^T \vec{\beta}}{\sqrt{\vec{c}^T (X^T X)^{-1} \vec{c} \frac{\mathrm{SSE}}{n-p}}} \sim t_{n-p}$

Proof. To show independence, it is sufficient to show that $\hat{\beta}$ and $(\hat{Q}^\perp)^T \vec{Y}$ are independent, where $\hat{Q}^\perp$ is defined as in (2.12). But they are jointly Gaussian, and

$$\mathrm{Cov}(\hat{\beta}(\vec{Y}), (\hat{Q}^\perp)^T \vec{Y}) = \sigma^2 (X^T X)^{-1} X^T \hat{Q}^\perp = 0$$

Notice that $(Q^\perp)^T X = 0$ (why?), so we have independence.

(2) is trivial. For (3), let $\vec{Z} = \vec{Y}/\sigma$. Then $\vec{Z} \sim \mathbf{N}_n(\vec{0}, I^{(n)})$. Then by Fischer Cochran Theorem, $\vec{Z}^T P \vec{Z}$ and $\vec{Z}^T (I - P) \vec{Z}$ are independent with

$$\vec{Z}^T P \vec{Z} \sim \chi^2_p$$
$$\mathrm{SSE}/\sigma^2 = \vec{Z}^T (I - P) \vec{Z} \sim \chi^2_{n-p}$$

Finally for (4), we have

$$T = \frac{\vec{c}^T\hat{\beta} - \vec{c}^T\vec{\beta}}{\sqrt{\vec{c}^T(X^TX)^{-1}\vec{c}\,\frac{\mathsf{SSE}}{n-p}}} = \frac{\frac{\vec{c}^T\hat{\beta}-\vec{c}^T\vec{\beta}}{\sqrt{\vec{c}^T(X^TX)^{-1}\vec{c}\,\sigma^2}}}{\sqrt{\frac{\frac{\mathsf{SSE}}{\sigma^2}}{n-p}}} \sim \frac{\mathbf{N}(0,1)}{\sqrt{\frac{\chi^2_{n-p}}{n-p}}} \sim t_{n-p}$$

□

We have therefore constructed a pivot for $\vec{c}^T\vec{\beta}$, and use it to construct confidence interval, perform hypothesis testing etc.