Lecture in Statistical Testing Theory:

HEINRICH MATZINGER
Georgia Tech
E-mail: matzi@math.gatech.edu

December 5, 2016

Contents

1 Basic definitions related to statistical testing for simple hypthesis

The main ideas and definitions related to statistical testing can be understood best when we think of a simple little guessing game involving several different dice. For this imagine the following game: Matzinger owns two four-sided dice. One has the same probabilities for each side and is hence a symmetric die. The other has an irregular shape with different probabilities for each side. Matzinger is going to throw one of the two dice. He will not tell you which die he has used. But Matzinger is going to tell you which number was obtained. This number will be denoted by X and you will have to guess which die has been used just based on the information which he gives you. Again, you know all the different probabilities for each die. We call the fact that the first die was used, the null-hypothesis and denote it by H_0. The fact that the second die has been used, will be called the *alternative hypothesis* and is denote by H_1. Trying to figure out which die was used based on one outcome X, is called *testing* H_0 *against the alternative hypothesis* H_1.

In general we do not assume that there is a probability to which die is used. Otherwise if there exists probabilities for the hypotheses H_0 and H_1, then we will say that we are doing Bayesian-statistics. The probabilities of the hypothesis, that is $P(H_0)$ and $P(H_1)$ are then called *prior probabilities*. Now assume that we would have a bag with 10 dies of the first type and 90 of the second type. And say, we would chose in that bag one of the dies at random so that each die has the same probability to be chosen. Then, we would be in the Bayesian statistics case, and $P(H_0) = 0.1$ and $P(H_1) = 0.9$. So, this would be a case where clearly it would make sense to assume the existence of prior probabilities. On the other hand assume that your are testing oranges for mercury at the CDC. The hypothesis would be that there is more than a certain legal among of Mercury in the oranges. Maybe suddenly the law changes and oranges can be imported from a country where before import was not possible. Then this might lead to some oranges with Mercury if in that country regulation is less strict. But, in this case we would not think of a prior probability of the oranges having Mercury, because that probability just change when the new situation occurred. .....

Now let us assume that the probabilities of the regular die are as follows:

$$P(X=1|H_0)=0.25, P(X=2|H_0)=0.25, P(X=3|H_0)=0.25, P(X=4|H_0)=0.25. \tag{1.1}$$

Assume that if the alternative hypothesis H_1 holds, that is if we use the skewed die, then the probabilities are as follows:

$$P(X=1|H_1)=0.4, P(X=2|H_1)=0.3, P(X=3|H_1)=0.2, P(X=4|H_1)=0.1. \tag{1.2}$$

We can summarize both probability models in one table

x	1	2	3	4
$P(X=x\|H_0)$	0.25	0.25	0.25	0.25
$P(X=x\|H_1)$	0.4	0.3	0.2	0.1

Then, your teacher Matzinger throws one of the two dice, but does not tell you which he used. Say he obtains the value 2, so that $X = 2$. Now, you have to guess based on that value $X = 2$ alone which of the two dice was used. Of course, you also know the two models under consideration. That is you know all the probabilities given in 1.1 and 1.2. There are different ways to make your decision as to which die was used based on the data X given to you.

One example of a test, would be a maximum-Likelihood based test. In this example, side 1 and 2 have bigger probability under H_1, than under H_0. On the other hand, the sides 2 and 3 have bigger probability under H_0. So, we define a first test which we call TEST!. The *decision rule* for **TEST1** would then be:

1. if side 1 or 2 come we reject the null-hypothesis H_0, that is the hypothesis that the symmetric die was used.
2. If side 3 or 4 appear, then we accept the null-hypothesis H_0, that is we accept the hypothesis that the die used was symmetric

Since we obtained a 2 in our example when throwing the die, we reject the H_0-hypothesis with our test TEST1. In other words, for definition a test, we just need to define an acceptance region. If then, the value X observed falls within the acceptance region, we accept H_0 whilst otherwise we reject it.

So for TEST1, we have $\{3, 4\}$ is the *acceptance* region. On the other hand the set $\{1, 2\}$ is the *rejection region* of TEST1 which is also called *critical region of the test.*

You will not know in general which die was used. Of course in our little game Matzinger could just tell you. But in real life, we usually will not know not for sure which model is the correct one, that is the one which generates the data. think for example X to be same data about nature which you have on a file. You may know two possible random models which could have generated the data X. But, nobody while tell you in the end which model was the correct one: you only have the data X to guess. So, we see, statistical hypothesis testing is about guessing which stochastic model is generating real life data, when you have the data available....

let us get back to the little example of the game played with Matzinger. There are two type of errors:

1. When the first die is used and we decide the reject (wrongly) that H_0-Hypothesis we say that we are making an *error of type I*. The probability of it given that the null-hypothesis holds, is called *significance of the test* and is denoted by α. Hence,

   $$\texttt{significance of test} = P(\text{We commit an error of type I}|H_0) = P(X \text{ is in rejection region}|H_0)$$

2. We call error of *type II*, an error which is committed whilst the alternative hypothesis hold. Thus, such an error consists in failing to reject H_0.(in our example: failing to reject H_0, despite as having used the second die). We denote the probability of an error of type II under the hypothesis H_1 by β and hence

   $$\beta = P(\text{We commit an error of type II}|H_1) = P(X \text{ is in acceptance region}|H_1)$$

furthermore, we call $1-\beta$ the *power of the test.*

Note, that we might never now for sure in the end what type of error we made, since we might never know for sure which die was used.

In our current example we find that the significance for TEST1, is equal to $\alpha = P(1|H_0) + P(2|H_0) = 0.25 + 0.25 = 0.5$. In our case, we have that $\beta = P(3|H_1) + P(4|H_1) = 0.2 + 0.1 = 0.3$ The probability of not making an error of type II under H_1 is 1 minus the probability of making such an error. Hence the power of a test is $1-\beta$. (CAREFUL, some books use β for the power instead of for the probability of a type II error. In our text-book β represents the power, but I use it for the probability of an error of type II).

We have thus two numbers which characteristic a test: the significance α and the power $1-\beta$. Typically we want the significance α to be small and the power β to be large. Usually there is a trade off between the two: if we need better significance, then we will construct another test, but which typically will have worse (less) power.

So, let us consider two possible tests foand compare them. Let $TEST1$ be the test we have considered so far and let $TEST2$ have acceptance region given by: $\{2,4\}$. So, schematically we have the following situation:

$TEST1$			a	a
$TEST2$		a		a
x	1	2	3	4
$P(X=x\|H_0)$	0.25	0.25	0.25	0.25
$P(X=x\|H_1)$	0.4	0.3	0.2	0.1

where a stands for acceptance region. Now let us compare the significance and power of the two tests under consideration in a table:

	α	$1-\beta$
$TEST1$	0.5	0.7
$TEST2$	0.5	0.6

We see both tests have significance 0.5. But TEST1 has bigger power than TEST2. So we can discard TEST2, since TEST1 clearly has only advantages. You can check that in the current example TEST1 among all possible tests with significance of 0.5 has most power. Such a test is called *most powerful* test. So, a most powerful test on a given significance level α is a test which maximizes the power among all tests with significance α in that given situation.

In many real life situations we will be given a requested significance level α and then be asked to find the most powerful test on that significance level. (Imagine again that you are testing oranges for Mercury at the CDC. Then, the significance is typically given by Law to you....) Now, say in the current example you would need a test with a significance of 5%. This in the current setting seems to not be possible, because the smallest probability you encounter is 25% under H_0. So, you would like to "split" one of these probabilities. The way to do so is by doing a *randomized test.* For example, when $X = 4$ you trow a five sided die with five equal likely sides: I,II,III,IV and V. We thus assume that

$$P(I) = P(II) = P(III) = P(IV) = P(V) = \frac{1}{5}$$

and that this die used is independent of the other dies. We could establish the following rule: When the data X is equal to 4 we throw that other die and if we get a I we reject the hypothesis. In all other cases we accept the hypothesis. The significance of this randomized test is then

$$P(\text{commit an error of type I}|H_0) = \frac{1}{5}P(X = 4|H_0) = \frac{1}{5}0.25 = 0.05$$

So, we get a 5% significance for this randomized test. Schematically we will represent a randomized test as follows: for each value of x we write the probability of accepting H_0 given that $X = x$ above the value of x. So, our randomized test can now be represented by

$P(we\ reject\ H_o\|X = x)$	0	0	0	0.2
x	1	2	3	4

In general the significance for a randomized test when we have a simple null-Hypothesis and a finite number of possibilities for values for X is

$$\texttt{significance of randomized test} = \sum_x P(rejection|X = x) \cdot P(X = x|H_0).$$

Another situation can be when we need randomization to construct a test with more power on a certain significance level. Assume for this slightly changed probabilities from our example:

$TEST1$			a	a
$TEST2$		a		a
x	1	2	3	4
$P(X = x\|H_0)$	0.25	0.25	0.251	0.249
$P(X = x\|H_1)$	0.4	0.3	0.2	0.1

So, we changed slightly, by no more than one percent the probability for 3 and 4. Everything, else remains the same. Now, in the situation before $TEST2$ could be discarded, because $TEST1$ had the same significance as $TEST2$ but a better power. In this new situation, if we do not allow randomized tests, we have that $TEST2$ becomes a most powerful test for its significance level of 0.501. (Most powerful among all non-randomized test with significance level 0.501). The reason is that $TEST2$ has significance equal to

$$P(1|H_0) + P(3|H_0) = 0.25 + 0.251 = 0.501$$

and $TEST1$ has significance 0.5. So the two tests can not be compared "officially". But in real life of course this is non-sense! The difference in significance of 0.01 is not going to matter, but the difference in power of 10% in important. So in real life we would most certainly prefer $TEST1$ over $TEST2$. The problem is that without randomizing we can not split those probabilities and hence a test which might in practice be non-sensical, can be the most powerful given its exact significance level. This is simply because it is impossible to put together another test with exactly the same significance. (there only a very limited number of possible tests....) Now, if we can randomize, then this is no longer a problem and a test like TEST2 is no longer optimal, because we can construct a randomized test with exactly the same significance but a better power. Here is what we do. We would like $TEST1$ instead of $TEST2$. But the problem is that $TEST1$ has a slightly different significance than $TEST2$. So, what we do is we slightly change $TEST1$ by randomization to obtain exactly the same significance than $TEST2$. But the randomized test will have a power close to $TEST1$ (though not exactly the same power) and hence beat $TEST2$ in power. To get the significane of 0.501 we will decide when we have a 1 or a 2 to reject the null-hypothesis. This gives only a significance of 0.5 as of yet. Now, we introduce the randomization: when we have $X = 3$, we flip a coin which is stongly biased to decide if we reject the hypothesis or not. That coin will have a probability of 1/251 to tell us to reject. With $X = 4$, we always accept H_0. The randomized test is defined in the table below by its conditional probability to reject H_0 given different values of X:

$P(\texttt{reject}\vert\mathtt{H_0},\mathtt{X}=\mathtt{x})$	1	1	$\frac{1}{251}$	0
x	1	2	3	4
$P(X = x\vert H_0)$	0.25	0.25	0.251	0.249
$P(X = x\vert H_1)$	0.4	0.3	0.2	0.1

And so the significance for this randomized test is now just equal to

$$\texttt{significance of randomized test} = \sum_x P(\texttt{reject}|\mathtt{H_0},\mathtt{X}=\mathtt{X}) \cdot \mathtt{P}(\mathtt{X}=\mathtt{x}|\mathtt{H_0}) =$$

$$= 1 \cdot 0.25 + 1 \cdot 0.25 + \frac{1}{251} \cdot 0.251 + 0 \cdot 0.249 = 0.501$$

which is exactly the same significance as $TEST2$. The power is

$$\texttt{power of randomized test} = \sum_{\mathsf{x}} \mathtt{P}(\texttt{reject}|\mathtt{H_1},\mathtt{X}=\mathtt{x})\mathtt{P}(\mathtt{X}=\mathtt{x}|\mathtt{H_1}) = \mathsf{1 \cdot 0.4 + 1 \cdot 0.3 + \frac{0.2}{251}} \approx \mathsf{0.701},$$

which is way better than the power $\beta = 0.6$ of TEST2!

2 Neyman-Pearson and how to find an optimal test with simple hypotheses

Let us define the *likelyhood ratio*:

$$ratio(x)\frac{P(X = x|H_0)}{P(X = x|H_1}$$

let us going back to the origin situation with two dice and let us record the Likelihood ratio in the table:

$TEST1$			a	a
$TEST2$		a		a
$ratio(x)$	0.625	$0.8\bar{3}$	1.25	2.5
x	1	2	3	4
$P(X = x \vert H_0)$	0.25	0.25	0.25	0.25
$P(X = x \vert H_1)$	0.4	0.3	0.2	0.1

We had seen that $TEST2$ is not optimal because in its acceptance region it contain 1 instead of 3. When we change the acceptance region of $TEST2$ by replacing 2 by 3, we improve the power whilst maintaining the significance at the same level. In other words, the reason why $TEST2$ is not optimal is that in the acceptance region we can replace a point with another point which has higher likelihood ratio. (In our case replacing in the acceptance region 2 by 3). Note that

$$ratio(2) = 0.8\bar{3} < ratio(3) = 1.25.$$

So, here the argument is that replacing a point in the acceptance region with another with higher ratio can only increase the power. (This argument works here because when changing 2 with 3, these points have the same probability under H_0. In most cases, this will not be the case.) So, in other words for our tests to be optimal they need the acceptance region to follow the order of increasing likelihood ratio. Test ! for example is optimal. the acceptance region corresponds to requesting that the ratio be bigger or equal to $0.8\bar{3}$. this corresponds to the points 3 and 4. In general it will not be the case, that under H_0 all points have the same probability. Non-the less it turns out in general when we test two simple hypothesis against one another that all optimal tests must follow the order of the likelihood ratio: the acceptance region can not contain a point with lower ratio than any point in the rejection area. At least if we allow of randomized tests. this is the content of the Neyman-Pearson Lemma below

Lemma 2.1 *Assume that we test the simple hypothesis H_0 against the simple alternative H_1. We allow for randomized tests. let*

$$ratio(x) := \frac{P(X = x|H_0)}{P(X = x|H_1)}$$

Then any optimal test for a given significance level satisfies:
there exists a non-random constant c so that:

1. *if $ratio(X) < c$ we reject*

2. *if $ratio(X) > c$ we accept*

On the other hand, any test which satisfies the two conditions above is Most powerful for its given confidence level.

So, when $ration(x) = c$ we can reject the hypothesis or accept or make randomized decision. We will always get a most powerful test. (Only we will change the significance depending on what we decide to do when $ratio(x) = x$.
We will give a detailed proof of this Lemma. The proof should be clear when we have finite number of possible numbers and under H_0 they all have the same probability. Now when this is not the case, but still assume finite many possibilities. Then you can split each of the possibilities into little sub-cases with equal probability.....more to come....

3 UMP-tests for composite alternatives

Next we are looking at a similar game but with 3 dice instead of only 3. So, we add an additional die call it hypothesis H_2. Again, I am going to throw one of these three dice and you will have to guess based on the result which die was used. Again for all three dice (i.e random models) we assume the probabilities known to you. But, I want you only to tell me if it is die 1 used or die number 2 or 3. If you think that it is die 2 or 3 which was used, I don't ask you to tell me which one of the two is most likely in your view. I only want you to reject the hypothesis H_0. So, in this case we taste the null-hypothesis H_0 against the alternative K which consists of two possibilities H_1 and H_2. Again, you will accept H_0 or reject it in favor of K. But if you reject in favor of K, we don't need you to tell us, which of the possibilities in K, that is H_1 or H_2 you find more likely. The hypothesis K is called *composite*, because it does not consist of only one model but of several. In real life we often test hypothesis against alternatives which are composite: for example test the null-hypothesis that there is no lead in water vs K being the alternative hypothesis that there is lead in the water. Then K is a composite hypothesis: there are many different non-zero possible level of lead in the water: Each such possibility is contained in K. Let us look at an example with three dice: the first dice is the null-hypothesis. The second and third dice constitute the alternative. We could for example have the following situation:

$TEST1$			a	a
$TEST2$		a		a
x	1	2	3	4
$P(X = x\|H_0)$	0.25	0.25	0.25	0.25
$P(X = x\|H_1)$	0.4	0.3	0.2	0.1
$P(X = x\|H_2)$	0.4	0.2	0.3	0.1

Now consider $test1$ it is optimal if we had only the second die, that is for H_1 and H_0. But, with the third die $TEST1$ is not optimal! For H_0 against the alternative H_2, it turns out $TEST2$ is better. This leads to that in this specific situation given above where the alternative K consists of H_1 and of H_2, there is not overall best test with significance 0.5. A test that is "overall best", that is that has maximum power for each alternative, is called *(UMP) Uniformly Most Powerful test.* By this we mean that for each H_i in K we have:

the test is a maximum power test. The reason why in the above situation there is not UMP-test, is that the likelihood rations for H_1 and H_2 do not lead to the same ordering and hence to the same optimal (most powerful) tests. Let us define the ratios:

$$ratio_1(x) = \frac{P(X = x|H_0)}{P(X = x|H_1)}, ratio_2(x) = \frac{P(X = x|H_0)}{P(X = x|H_2)}.$$

Let us show these ratios in our table

$TEST1$			a	a
$TEST2$		a		a
$ratio_2(x)$	0.625	1.25	$0.8\bar{3}$	2.5
$ratio_1(x)$	0.625	$0.8\bar{3}$	1.25	2.5
x	1	2	3	4
$P(X = x\|H_0)$	0.25	0.25	0.25	0.25
$P(X = x\|H_1)$	0.4	0.3	0.2	0.1

So, $TEST1$ is better for H_1 because $ratio_1(2) < ratio_1(3)$. On the opposite, $TEST2$ is better with hypothesis H_2 because $ratio_2(2) > ratio_2(3)$. So our problem of not being able to find a UMP tests with significance 0.5 in the current situation comes from " the ordering induced by $ratio_1$ and $ratio_2$ being different on the points 2 and 3. When this is not the case, we can always find UMP-test at all significance levels: as matter of fact for each H_1 and H_2 separately the best rejection regions will coincide. This is the content of the next lemma

Lemma 3.1 *Assume that we have only a finite number of values for X to take. Denote the finite set of these values by Ω. We assume that we test a simple hypothesis H_0 against the composite alternative hypothesis K. We allow for randomized tests. Then a necessary and sufficient condition for the existence of UMP-tests at all significance levels is that the order induced on Ω by the different ratios of the models in K are all compatible. More precisely, the condition is that for any $x, y \in \Omega$, we have that for any two models H_1 and H_2 in K, the following holds:*
if

$$ratio_1(x) > ratio_1(y)$$

then

$$ratio_2(x) \geq ratio_2(y)$$

where for all $x \in \Omega$

$$ratio_1(x) := \frac{P(X = x|H_0)}{P(X = x|H_1)}, ratio_2(x) = \frac{P(X = x|H_0)}{P(X = x|H_2)}.$$

It is easy often easiest to verify if all the ratios for models inside the alternative hypothesis induce the same order on Ω, when there is a test statistic $T(x)$ so that all the ratios are monotone functions of $T(.)$. In that case we get the following result which also holds when we have models defined with densities rather than on finite probability space:

Theorem 3.1 *Assume that we are testing a simple null-hypothesis H_0 against a composite hypothesis K. We allow for randomized tests. Assume that there exists a test statistic $T(x)$ so that for every model H_i in K the likelihood ratio of H_i is an increasing function of $T(x)$. That means that if we define the likelihood ratio i: $ration_i(x) := \frac{P(X=x|H_0)}{P(X=x|H_i)}$ then we request that this ratio can be written as an increasing function of $T(x)$. Hence there should exist a (non-random) increasing function h_i so that $ratio_i(x) = h_i(T(x))$ for all $x \in \Omega$. Under that condition any test for which there exists a constant c so that*

1. *If $T(X) > c$ we a.s. accept H_0.*
2. *If $T(X) < c$ we a.s. reject H_0*

is a UMP-test. On top of this any UMP-test must satisfy the above two conditions.

Definition 3.1 *A type of family of random models where it is particularly easy to find a test statistic $T(x)$ for which all the models in the alternative have likelihood ratio which is monotone in $T(X)$ (so that the above theorem applies) is the called exponential model. The definition goes as follows: we call a family of probability models exponential family iff any probability model inside the family can be written in the form*

$$P(X = x|H_\theta) = h(x)g(\theta)\exp(T(x) \cdot \nu(\theta)).$$

Now, there

Note that in the case of an exponential family, the likely hood ratio when the null hypothesis and a H_1 are both from the exponential family, then the likelihood ratio is equal to

$$ratio_1(x) = \frac{P(X = x|H_0)}{P(X = x|H_1)} = \frac{h(x)g(\theta_0)\exp(T(x) \cdot \nu(\theta_0)}{h(x)g(\theta_1)\exp(T(x) \cdot \nu(\theta_1)} = \frac{g(\theta_0)}{g(\theta_1)}\exp(T(x)\cdot(\nu(\theta_0)-\nu(\theta_1))).$$

So we see immediately that if the null-hypothesis and all the models in the alternative-hypothesis K are in an exponential family then all the likelihood ratios are monotone in $T(x)$. (Here the $T(x)$ of the definition of the exponential family). So, then it is easy to get thanks to the theorem above UMP tests based on the statistic $T(X)$. These test will be of the type $T(X) > c$ accept if $T(X) < c$ and if $T(X) = c$ do whatever you want (what is convenient to get the right significance, for example randomize). for this to work, we simply need that for all parameters θ_0 in the null-hypothesis and all θ_1 in the alternate hypothesis, we have

$$\nu(\theta_0) - \nu(\theta_1) > 0.$$

4 Unbiased tests

Let us assume that X is a random variable with probability given by

$$c(\theta) \cdot e^{\theta \cdot T(x)} h(x),$$

where θ is the parameter. So, X is a variable from an exponential family. Let θ_0 be a non-random value of the parameter. We have seen in the homework that typically with exponential family we can find optimal tests for a hypothesis $H_0 : \theta = \theta_0$ against $K_1 : \theta > \theta_0$. Similarly we could find optimal tests for testing $H_0 : \theta = \theta_0$ against $K_2 : \theta < \theta_0$. But we can not find an optimal test (UMP test) for testing $H_0 : \theta = \theta_0$ against $\theta \neq \theta_0$. The reason is that for testing H_0 against $\theta < \theta_0$ we find a UMP test which is just the opposite from testing H_0 against $\theta > \theta_0$. To see why this is true, simply remember that by Neymann-Pearson the optimal tests (UMP) are obtained by putting

$$ratio_0(X) \geq \texttt{constant}$$

Now, the ratio is

$$\frac{P(X = x|\theta_0)}{P(X = x|\theta_1)} = e^{(\theta_0 - \theta_1)T(X)}.$$

So putting the expression on the right side of the last equation above bigger than a constant cst we find if $\theta_0 - \theta_1 > 0$ the following:

$$T(X) \geq \frac{\ln(cst)}{\theta_0 - \theta_1} \tag{4.1}$$

whilst if $\theta_0 - \theta_1 < 0$ then we get the opposite, because multiplying an inequality by a negative, inverses the inequality:

$$T(X) \leq \frac{\ln(cst)}{\theta_0 - \theta_1}. \tag{4.2}$$

Note that when we test $H_0 : \theta = \theta_0$ against $K : \theta = \theta_1$, then putting the ratio $P(X = x|\theta_0)/P(X = x|\theta_1)$ bigger than a constant yields the acceptance region for the optimal tests. In other words, the acceptance region for the optimal tests would be given by 5.12 if we test against the alternative $\theta < \theta_0$, but the acceptance region would be given by 5.13 if we test against $\theta > \theta_0$. So, there is no UMP-test valid for both alternatives: when the parameter in the alternative is above and when it is below θ_0.

There is one way around this: we are going to introduce the concept of *unbiased* tests. We will then restrict our attention to unbiased tests. And we will see that in a situation like the above (more precisely as soon as we have an exponential family with a one dimensional parameter) that if we do not consider all tests but only unbiased ones, there will be optimal tests among the unbiased ones. Optimal in the sense that on a given significance level they have maximum power for each value of the parameter in the alternative. Such tests will then be called UMP-unbiased tests. We will show that for any exponential family and for the a hypothesis $H_0 : \theta \in [a, b]$ against the alternative $K : \theta \notin [a, b]$ there always exists a UMP-unbiased test on any significance level. Furthermore these UMP-unbiased tests (for exponential families only) are of the type: we have an interval $[c_1, c_2]$. if $T(X)$ is strictly within the interval we accept H_0. If $T(X)$ is strictly outside we reject H_0. If $T(X)$ is on the border, then we might need to randomize the test. But randomization is never needed if we work with probability densities.

OK, now that we announced what we are going to do, let us explain first the concept of unbiased test. Let us start with an example:

Let us assume that X is a normal variable with $\sigma = 1$ and μ not known. Say we want to test $H_0 : \mu = 25$

against the alternative $K : \mu \neq 25$. This situation could happen for example, when we work in a physics lab. Then, μ could be the speed of a particle which we like to determine. We measure the speed, but as usual we make a small measurement error ϵ which is supposed to be random. So

$$X = \mu + \epsilon$$

The speed μ of the particle is not known, but is not random. It is a constant of physics which we try to determine. We get

$$E[X] = E[\mu + \epsilon] = \mu + E[\epsilon] = \mu + 0 = \mu,$$

where we assumed the expected measurement error to be 0.

Anyhow, so maybe a theoretical paper was published where some theoretical calculation lead to a scientist postulating that the speed should be 25. Now, we are going to verify this making a measurement in a lab and of course, as mentioned, there is always going to be a small measurement error. We assume the error to have a normal distribution. The average size of the error be $\sigma = 1$. Let us consider several possible tests and then think which ones make more sense. First we could consider a test $TEST1$ so that when X is between 23 and 27 we accept the hypothesis $H_0 : \mu = 25$ and reject H_0 when X is outside that interval. Another test $TEST2$ could be defined as follows: accept $H_0 : \theta = 25$ when X is between 24 and 28, and reject H_0 otherwise.

Well it should be pretty clear in the current situation with the task at hand that we would prefer $TEST1$ over $TEST2$. But what is wrong with using $TEST2$ for testing $H_0 : \mu = 25$ against $K : \mu \neq 25$? Well the problem with $TEST2$ is that the probability to be in the acceptance region $[24, 28]$ is much higher for $\theta = 26$ (by symmetry in this case) rather than for our hypothesis $H_0 : \theta = 25$:

it may make sense to use $TEST2$ for testing $\mu = 26$ as null hypothesis but not for our null-hypothesis $\mu = 25$. So formally, the reason why we do not want $TEST2$ for testing $\mu = 25$, (but it would be OK for testing $\mu = 26$) is that

$$P(X \in [24, 28]|\mu = 25) < P(X \in [24, 28]|\mu = 26).$$

The above inequality is the same as the following:

$$P(TEST\ 2\ accepts\ H_o|\mu = 25) < P(TEST\ 2\ accepts\ H_o|\mu = 26).$$

We see that with $TEST2$ if the parameter is $\theta = 26$, we have a greater probability to fall into the acceptance region then for $\theta = 25$. And this is precisely what we don't want. So, we want to avoid such a situation where under a false parameter the acceptance probability of H_0 is higher than if we have the parameter corresponding to H_0 in our case $\mu = 25$. A test which does not have the same problem as $TEST2$ has with our $H_0 : \mu = 25$ is called *unbiased* test. In our current example $TEST1$ is unbiased for $H_0 : \mu = 25$ against $K : \mu \neq 25$, but $TEST2$ is not unbiased for $H_0 : \mu = 25$ against $K : \mu \neq 25$. So, an unbiased test here in this example with $H_0 : \mu = 25$ would be any test so that:

the probability of acceptance of H_0 when $\mu = 25$ is bigger or equal to the probability of acceptance for any $\mu \neq 25$. So for example if $TEST3$ would accept $H_0 : \mu = 25$ if $X \in [a, b]$ and reject otherwise, then for $TEST3$ to be unbiased we would simply request that for any $\mu_1 \neq 25$ we have:

$$P(X \in [a, b]|\mu = 25) \geq P(X \in [a, b]|\mu_1).$$

So, for testing a simple hypothesis $H_0 : \mu = \mu_0$ against an alternative $H_0 : \mu \neq \mu_0$, saying that a test TEST is unbiased simply means that the probability of acceptance of the null-hypothesis is maximal for $\mu = \mu_0$. In other words $TEST$ is unbiased iff

$$P(\text{TEST accepts } H_0|\mu) = P(\text{X is in acceptance region of TEST}|\mu)$$

is maximal for $\mu = \mu_0$.

Now, we need to also define unbiasedness when the H_0- hypothesis is not simple but say is of the type "parameter μ in parameter-set I_0 against $K : \mu \in I_1$". (Of course we assume I_0 and I_1 have an empty intersection). In that case, we say that a test $TEST$ is *unbiased* for $H_o : \mu \in I_0$ against $K : \mu \in I_1$, iff for all $\mu_0 \in I_0$ and all $\mu_1 \in I_1$ we have

$$P(\text{TEST accepts } H_0|\mu_0) \geq P(\text{TEST accepts } H_0|\mu_1)$$

or equivalently:

$$P(\text{X is in acceptance region of TEST}|\mu_0) \geq P(\text{X is in acceptance region of TEST}|\mu_1).$$

Now, we will not just be interested in unbiased tests, but we will want to determine the most powerful tests among such tests. As usual we fix a significance level $\alpha \in [0, 1]$ and then look if there is among all unbiased tests with that given significance one which is most powerful for any $\mu_1 \in I_1$. So formally a UMP-unbiased test with significance α is an unbiased test $TEST$ with significance α so that for any $\mu_1 \in I_1$ we have:
for any unbiased test $TEST'$ with significance α, we have

$$P(\text{TEST accepts } H_0|\mu = \mu_1) \leq P(\text{TEST' accepts } H_0|\mu = \mu_1). \tag{4.3}$$

To understand the above inequality note that one minus the probability of acceptance of H_0 is the power. Now, it is always easy to find an unbiased test $TEST$ which would satisfy inequality 4.3 for one given value of the parameter $\mu_1 \in I_1$. the problem is to find a unbiased test $TEST$ which satisfies inequality 4.3 for all $\mu_1 \in I_i$ at the same time! Usually this is not possible, but for exponential families with a one dimensional parameter μ it will always be possible: thanks to the following theorem:

Theorem 4.1 *Let us assume that X is a random variable with probability given by*

$$c(\theta) \cdot e^{\theta \cdot T(x)} h(x),$$

where θ is the parameter. So, X is a variable from an exponential family. Let I_0 and I_1 be two disjoint subsets of the parameter set. Let $\alpha \in [0, 1]$. There exists a UMP-unbiased test on the significance level α for testing $H_0 : \theta \in I_0$ against $H_1 : \theta \in I_1$. Furthermore, that UMP-test can be chosen like accept inside a certain interval and reject outside whilst on the border you may have to randomize. More precisely, there exists $c_1 < c_2$ and $\gamma_1, \gamma_2 \in [0, 1]$, so that the following constitutes a UMP-unbiased test at level α:

- *accept if $c_1 < T(X) < c_2$*
- *reject if $T(X) > c_2$ or $T(X) < c_1$*
- *if $T(X) = c_1$ accept with probability γ_1*
- *if $T(X) = c_2$ accept with probability γ_2*

The randomization on the border is not necessary if we deal with density functions rather than discrete probabilities.

Now, the question maybe for a given hypothesis $H_0 : \theta \in I_0$ and a given significance level α how do we determine a UMP-unbiased test? Of course we assume that the variable at hand, that is X is from a one-dimensional exponential family since otherwise there might not be such a test. Typically we will have for I_θ and interval $[a, b]$ which may contain a single point that is $[a, a]$ and we will test against $\mu \notin [a, b]$ or against $\mu \neq a$. In almost all real life situations, the function

$$\mu \mapsto P(X \in \text{ acceptance region}|\mu)$$

will be a continuous function. so, then if a, b are border points (meaning that there values as close as we want from both withing I_0 and within I_1), then a necessary condition (not sufficient though) for the test being unbiased with significance α is that

$$P(X \in \text{ acceptance region}|\mu = a) = P(X \in \text{ acceptance region}|\mu = b) = 1 - \alpha.$$

To see why this holds, simply imagine that one of the two probabilities in the last equation above would be strictly bigger than the other. Say for example, the probability above with $\mu = a$ is strictly bigger than for $\mu = b$. Then, we could find a point μ_1 in the parameter space very close to a but still in I_1 so that

$$P(X \in \text{ acceptance region}|\mu = \mu_1)$$

is as close as we want to

$$P(X \in \text{ acceptance region}|\mu = a)$$

by continuity. So, by taking μ_1 close enough to a we could get its acceptance probability strictly bigger than for $\mu = b$:

$$P(X \in \text{ acceptance region}|\mu = \mu_1) > P(X \in \text{ acceptance region}|\mu = b)$$

and that would imply that the test is not unbiased since $\mu_1 \in I_1$.
Anyhow, so we have seen that we have a necessary condition for a test to be unbiased: on the border points of I_0 the value of the acceptance probability must everywhere be equal to $1 - \alpha$. (Assuming the acceptance probability to be continuous in the parameter which is almost always the case in practice). But now we have a condition for unbiasedness, but what we need is not just unbaisedness but a UMP-unbiased. So, how are we going to find a UMP-unbiased? (Again this only works for exponential families, since otherwise UMP-unbiased in most cases does not even exist). Well we are going to use our Theorem which guaranties the existence of a UMP-unbiased test (for exponential families with a density) of the type: accept inside $[c_1, c_2]$ and reject if outside. So, for the significance level α, what we mentioned before implies that for testing $H_0 : \mu \in [a, b]$ against $K : \mu \notin [a, b]$

with a UMP-test of the type: “accept when in $[c_1, c_2]$ and reject otherwise, we need to have

$$P(X \in [c_1, c_2]|\theta = a) = P(X \in [c_1, c_2]|\theta = b) = 1 - \alpha. \tag{4.4}$$

Now, the above equation in principle does not guaranty a UMP-unbiasedness since it is only a necessary condition for unbiasedness. But in practice it will in most cases give the UMP-unbiased test at level α. (Again only for exponential family). why? because in most applied situation we will encounter in real life, there will be a unique solution c_1, c_2 to the equation 5.3. This then implies that we got the UMP-unbiased test, since the UMP-unbiased test satisfies 5.3. So, in practice, when dealing with exponential family with one parameter, we usually simply solve equation 5.3, for the unknown c_1 and c_2 and when there is a unique solution, this then implies that we have found the UMP-unbiased test at level α.

When dealing with an exponential family with discrete probabilities space, things are a little more complicated since you have to determine also γ_1 and γ_2. So, then what you do is for every pair of numbers $c_1 \leq c_2$ in your discrete space you try to determine γ_1 and γ_2 so that equation 5.3 holds. You keep on searching until you find a pair c_1, c_2 for which there exists γ_1 and γ_2 so that equation 5.3 is satisfied. This is a little more work of course.... So for example if X can take any of the values $\{1, 2, 3, 4\}$. then you have to try for any $c_1 \leq c_2$ with $c_1, c_2 \in \{1, 2, 3, 4\}$. This is 4! possibilites that is 24 couples c_1, c_2 that is

$$(1, 1), (1, 2), (1, 3), (1, 4), (2, 2), (2, 3), \ldots$$

5 Mutlidimensional linear regression and other linear models

First assume that we have a data with some cows and how much milk they produce. The data-spread sheet could look as follows for example:

cow nb	*milk*	*weigth*	*age*
1	4	2	5
2	5	3	6
3	8	4	5
4	9	4	4

So, if we look carefully at this data set, we will see that there is a simple “formula” for finding the amount of milk produced by each cow based on age and weight:

$$milk = 5 + 2 \times weigth - age \tag{5.1}$$

and it works for all the cows in the current data-set. Now such a relationship between milk and weight and age might be useful for predicting how much a cow which we have not bougth yet will produce in milk: maybe we are told the age and the weight, but not the amount of milk produced. Again, here we have that for this data set the relationship

5.1 works exactly. In general we will not be able to find such a linear relationship with fits the data exactly. Let us explain. Say, we are looking for coefficients α, β_1 and β_2 so that

$$milk = \alpha + \beta_1 \times weigth + \beta_2 \times age \tag{5.2}$$

holds for each cow. The typical real life situation will be hundred of cows in a data-set and not just 4 like here. So, with hundred of equations (one equation for each cow), but only 3 variables, we typically get a system of linear equations which is over-determined, that is has no solution. So, when trying to find a type of relationship between the milk and age and weight like given in 5.2, we will have in general to content ourselves with finding parameters α, β_1 and β_2 which comes as close as is possible to 5.2 in the data-set available. For, this consider for example the following data-set:

cow nb	*milk*	*weigth*	*age*
1	3.5	2	5
2	4.5	3	6
3	7.5	4	5
4	8.5	4	4

In this current example now, there is no parameters α, β_1 and β_2 which make 5.2 hold exactly for each cow. So, instead we look for parameters which get "as close as possible to making 5.2 hold for each cow". That is for given α, β_1 and β_2, we take the square of the difference between that actual amount of milk produced by the cow and what our formal with the coefficients would give. Then, we minimize the sum of the squares of these "approximation errors" in our data. So, we the current data-set, we minimize

$$(3.5-\alpha-\beta_1 2-\beta_2 5)^2+(4.5-\alpha-\beta_1 3-\beta_2 6)^2+(7.5-\alpha-\beta_1 4-\beta_2 5)^2+(8.5-\alpha-\beta_1 4-\beta_2 4)^2$$

You can now minimize the above expression by finding the partial derivatives according to α, then according to β_1 and according to β_2. Setting each of these three partial derivative s equal to 0 yields a system of three linear equations. When you solve this system, you get the coefficients α, β_1 and β_2 which come closest to predicting exactly what the milk for each cow is as a linear function of age and weight. Let us do this, but in a general context with letters instead of numbers. In this way, we will have a generally valid formula. So, let y_i denote the amount of milk produced by the i-th cow. Let x_i^{age} denote her age and let x_i^{weigth} denote the wiegth of the i-th cow. Then, we want to minimize the following function:

$$SS = \sum_{i=1}^{n}(y_i - \alpha - \beta_1 x_i^{weigth} - \beta_2 x_i^{age})^2$$

where n denotes the number of cows and the y_i, x_i^{age} and x_i^{weigth} have to be thought of as real numbers known to us from our data-spread sheet. To minimize SS we find the partial derivatives according to α, β_1 and β_2 and set them equal to 0. In this manner we find three linear equations:

$$\frac{dSS}{d\alpha} = -2\sum_{i=1}^{n}(y_i - \alpha - \beta_1 x_i^{weigth} - \beta_2 x_i^{age}) = 0 \tag{5.3}$$

$$\frac{dSS}{d\beta_1} = -2\sum_{i=1}^{n} x_i^{weigth}(y_i - \alpha - \beta_1 x_i^{weigth} - \beta_2 x_i^{age}) = 0 \tag{5.4}$$

$$\frac{dSS}{d\beta_2} = -2\sum_{i=1}^{n} x_i^{age}(y_i - \alpha - \beta_1 x_i^{weigth} - \beta_2 x_i^{age}) = 0 \tag{5.5}$$

the value for α, β_1 and β_2 which satisfy the three liner equations above, that is 5.3, 5.4 and 5.5 are "our estimate of what the best way would be to approximate the milk produced by a cow using linear expression of age and weight". To denote estimates in statistics one usually uses a hat, so we will denote by

$$\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2$$

the coefficients α, β_1 and β_2 which satisfy 5.3,5.4,5.5. (note in general there will be exactly one solution, since we have the same number of equations as we have unknowns). In other words, $\hat{\alpha}$, $\hat{\beta}_1$ and $\hat{\beta}_2$ are defined by the three following equations:

$$\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}_1 x_i^{weigth} - \hat{\beta}_2 x_i^{age}) = 0 \tag{5.6}$$

$$\sum_{i=1}^{n} x_i^{weigth}(y_i - \hat{\alpha} - \hat{\beta}_1 x_i^{weigth} - \hat{\beta}_2 x_i^{age}) = 0 \tag{5.7}$$

$$\sum_{i=1}^{n} x_i^{age}(y_i - \hat{\alpha} - \hat{\beta}_1 x_i^{weigth} - \hat{\beta}_2 x_i^{age}) = 0 \tag{5.8}$$

which in vector notation using the dot-product are equivalent to

$$\vec{1} \cdot \left(\vec{y} - \hat{\alpha}\vec{1} - \hat{\beta}_1 \vec{x}^{weigth} - \hat{\beta}_2 \vec{x}^{age}\right) = \vec{0} \tag{5.9}$$

$$\vec{x}^{weigth} \cdot \left(\vec{y} - \hat{\alpha}\vec{1} - \hat{\beta}_1 \vec{x}^{weigth} - \hat{\beta}_2 \vec{x}^{age}\right) = \vec{0} \tag{5.10}$$

$$\vec{x}^{age} \cdot \left(\vec{y} - \hat{\alpha}\vec{1} - \hat{\beta}_1 \vec{x}^{weigth} - \hat{\beta}_2 \vec{x}^{age}\right) = \vec{0} \tag{5.11}$$

5.1 The statistical model and the main properties of the estimates

Next we want to introduce a rigorous statistical model and show that the "estimates" $\hat{\alpha}$, $\hat{\beta}_1$, $\hat{\beta}_2$ make sense in a statistical formal sense. Our model is now the following: we assume that for each cow, the amount of milk produced Y_i is a random variable, which is equal to $\alpha + \beta_1 x_i^{weigth} + \beta_2 x_i^{age}$ plus a random term denoted by ϵ_i. So,

$$\text{milk of cow number } i = Y_i = \alpha + \beta_1 x_i^{weigth} + \beta_2 x_i^{age} + \epsilon_i$$

We assume the terms ϵ_i independent of each other and having expectation 0. At first we will also assume that the terms ϵ_i are normal $\mathcal{N}(0, \sigma)$ where σ is identical for each cow. So, this is as if nature would determine how much milk a cow produces by first computing the number $\alpha + \beta_1 \cdot weigth_i + \beta_2 \cdot age_i$ and then adding the random term ϵ_i. That random term is "the individual factor" of the cow, that is how much the cow fluctuates from what the formula would predict. Also, because we assume $E[\epsilon_i] = 0$ we find

$$E[Y_i|age, weigth] = \alpha + \beta_1 \cdot age + \beta_2 \cdot weigth + E[\epsilon_i] = \alpha + \beta_1 \cdot age + \beta_2 \cdot weigth.$$

So, the coefficients α, β_1 and β_2 are non-random and are the same for each cow. But we don't know them: only nature knows them. So, we hope that our estimates $\hat{\alpha}$, $\hat{\beta}_1$ and $\hat{\beta}_2$ will be close to the true (but unknown) values for these parameters. We will see below that indeed this is true when we have enough cows. But, first let us examine the property of these estimators. We find the following list of properties:

1. The estimates are unbiased that is:

$$E[\hat{\alpha}] = \alpha, E[\hat{\beta}_1] = \beta_1, E[\hat{\beta}_2] = \beta_2.$$

2. The estimates are maximum-likelihood estimates.

3. The model is an exponential family with a multidimensional parameter θ.

4. since the $T(.)$ statistic in multidimensional exponential families are complete as soon as the parameters are defined on a large enough set (in our case there is no restriction on the parameters), we get that the family is complete and hence there is only one unbiased estimator which depends only on $T(.)$ So, if we want unbiased estimators, $\hat{\alpha}$, $\hat{\beta}_1$ and $\hat{\beta}_2$ is the only possibility which makes sense. The $\vec{T}(.)$ statistic is sufficient in a multidimensional exponential family. so, one can throw out everything else. But the $\vec{T}(.)$ statistics is complete that means an unbiased estimator which depends only on $\vec{T}$ is unique. Another way, to express this is that our estimators are, among all unbiased estimators the only ones with lowest variance.

Proof of unbiased first note that the expected amount of milk of the i-th cow is

$$\begin{aligned}
&E[Y_i] = E[\alpha + \beta_1 x_i^{weight} + \beta_2 x_i^{age} + \epsilon_i] = \\
&E[\alpha] + E[\beta_1 x_i^{weight}] + E[\beta_2 x_i^{age}] + E[\epsilon_i] = \\
&\alpha + \beta_1 x_i^{weight} + \beta_2 x_i^{age}
\end{aligned}$$

Where we used the fact the expectation of a constant is the constant itself and we also used that the terms ϵ_i have 0 expectation. We will now use the expression we found for $E[Y_i]$ by taking the expectation on both sides of 5.6, 5.7 and 5.8. We find the following three equations

$$\begin{aligned}
\sum_{i=1}^{n}(E[Y_i] - E[\hat{\alpha}] - E[\hat{\beta}_1]x_i^{weight} - E[\hat{\beta}_2]x_i^{age}) =& 0 \\
\sum_{i=1}^{n} x_i^{weight}\left(E[Y_i] - E[\hat{\alpha}] - E[\hat{\beta}_1]x_i^{weight} - E[\hat{\beta}_2]x_i^{age}\right) =& 0 \\
\sum_{i=1}^{n} x_i^{age}\left(E[Y_i] - E[\hat{\alpha}] - E[\hat{\beta}_1]x_i^{weight} - E[\hat{\beta}_2]x_i^{age}\right) =& 0.
\end{aligned}$$

When, in the three equations above, we replace $E[Y_i]$ by $\alpha + \beta_1 x_i^{weigth} + \beta_2 x_i^{age}$ then we get the following three equations:

$$\begin{aligned}
\sum_{i=1}^{n}\left[(\alpha - E[\hat{\alpha}]) + (\beta_1 - E[\hat{\beta}_1])x_i^{weight} + (\beta_2 - E[\hat{\beta}_2)])x_i^{age})\right] &= 0 \\
\sum_{i=1}^{n} x_i^{weight}\left[(\alpha - E[\hat{\alpha}]) + (\beta_1 - E[\hat{\beta}_1])x_i^{weight} + (\beta_2 - E[\hat{\beta}_2])x_i^{age}\right] &= 0 \\
\sum_{i=1}^{n} x_i^{age}\left[(\alpha - E[\hat{\alpha}]) + (\beta_1 - E[\hat{\beta}_1])x_i^{weight} + (\beta_2 - E[\hat{\beta}_2])x_i^{age}\right] &= 0
\end{aligned}$$

Which in vector form can be written as

$$0 = \Big((\alpha - E[\hat{\alpha}])\vec{\mathbf{1}} + (\beta_1 - E[\hat{\beta}_1])\vec{x}^{weight} + (\beta_2 - E[\beta_2])\vec{x}^{age}\Big) \cdot \vec{\mathbf{1}} \tag{5.12}$$

$$0 = \Big((\alpha - E[\hat{\alpha}]) \cdot \vec{\mathbf{1}} + (\beta_1 - E[\hat{\beta}_1])\vec{x}^{weight} + (\beta_2 - E[\hat{\beta}_2])\vec{x}^{age}\Big) \cdot \vec{\mathbf{x}}^{weight} \tag{5.13}$$

$$0 = \Big((\alpha - E[\hat{\alpha}]) \cdot \vec{\mathbf{1}} + (\beta_1 - E[\hat{\beta}_1])\vec{x}^{weight} + (\beta_2 - E[\hat{\beta}_2])\vec{x}^{age}\Big) \cdot \vec{\mathbf{x}}^{age} \tag{5.14}$$

where $\vec{x}^{weigth}$ refers to the vector of the weights of the cows:

$$\vec{x}^{weigth} := (x_1^{weigth}, x_2^{weigth}, \dots, x_n^{weigth})^t$$

and $\vec{x}^{age}$ is the vector of the ages:

$$\vec{x}^{age} := (x_1^{age}, x_2^{age}, \dots, x_n^{age})^t.$$

Furthermore, the column vector of length n and all entries equal to 1 is denoted by $\vec{\mathbf{1}}$. We can now add the three equations 5.12, 5.13 and 5.14 after multiplying the first by $\alpha - E[al\hat{p}ha]$, the second by $\beta_1 - E[\hat{\beta}_1]$ and the third by $\beta_2 - E[\hat{\beta}_2]$. This then leads to

$$0 = \Big((\alpha - E[\hat{\alpha}]) \cdot \vec{\mathbf{1}} + (\beta_1 - E[\hat{\beta}_1])\vec{x}^{weigth} + (\beta_2 - E[\hat{\beta}_2])\vec{x}^{age}\Big)^2,$$

which implies that

$$\vec{0} = (\alpha - E[\hat{\alpha}]) \cdot \vec{\mathbf{1}} + (\beta_1 - E[\hat{\beta}_1])\vec{x}^{weigth} + (\beta_2 - E[\hat{\beta}_2])\vec{x}^{age} \tag{5.15}$$

Now we will assume that $\vec{\mathbf{1}}$, $\vec{x}^{weigth}$ and $\vec{x}^{age}$ are linearly independent. Under that assumption, equation 5.15, implies

$$E[\hat{\alpha}] = \alpha$$
$$E[\hat{\beta}_1] = \beta_1$$
$$E[\hat{\beta}_2] = \beta_2$$

and hence our estimators are unbiased.

Proof that our estimator is a Mixmum-Likelihood estimator

We first assume that $\sigma = \sqrt{VAR[\epsilon_i]}$ is known. This could be the case if for example we work in a lab and are trying to determine the size of an object depending on temperature and pressure. Then, we measure the size for different values of the pressure and temperature. There could be a linear relationship:

$$size = \alpha + \beta_1 \cdot temp + \beta_2 \cdot pressure.$$

But, when we measure we always make small measurement errors: so, let Y_i be our i-th measurement of size. We assume that we had a given temperature t_i and a given pressure p_i which are known to us. So, the true size for the i-the measurement is

$$size_i = \alpha + \beta_1 \cdot t_i + \beta_2 \cdot p_i.$$

But the i measurement is the true size plus a random error denoted by ϵ_i. So, if Y_i denotes our i-th measurement, we have

$$Y_i = \alpha + \beta_1 t_i + \beta_2 p_i + \epsilon_i,$$

where ϵ_i is the i-th measurement error. We assume again all measurement errors to have expectation 0 and be independent of each other. the precision of your measurement error is given by the standard deviation $\sigma = \sigma_{\epsilon_i}$. We assume all the measurement errors to have identical standard deviation. If you know, your measurement tools well, because you work often with them, then you will know the average size of your measurement error, thus you will know σ. All this to justify why we can sometimes consider σ known.....

So, let us assume σ is known. Let us go back to our cows for this is a terminology which we already know. Now, we assume the random measurement errors ϵ_i to be normal $\mathcal{N}(0, \sigma)$. Remember our main equation

$$Y_i = \alpha + \beta x_i^{weight} + \beta_2 x_i^{age} + \epsilon_i.$$

Here $\alpha + \beta x_i^{weight} + \beta_2 x_i^{age}$ is a constant and not random. When we add a non-random number to a normal, we get again a normal. So, Y_i is a normal with expectation given by

$\mu_i := \alpha + \beta x_i^{weight} + \beta_2 x_i^{age}$ and standard deviation σ. The probability density function of Y_i is thus up to a constant equal to $exp(-(y_i - \mu_i)^2/2\sigma^2)$. For independent variables the joint density is the product of their individual densities. We assumed the measurement errors ϵ_i to be independent of each others. Hence, the joint density of our data set

$$(Y_1, Y_2, \ldots, Y_n)$$

is equal to

$$P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n | \alpha, \beta_1, \beta_2) =$$
$$= \frac{1}{\sigma^n \sqrt{(2\pi)^n}} \exp(-(y_1 - \mu_1)^2/2\sigma^2) \cdot \exp(-(y_2 - \mu_2)^2/2\sigma^2) \cdot \ldots \cdot \exp(-(y_n - \mu_n)^2/2\sigma^2) =$$
$$= \frac{1}{\sigma^n \sqrt{(2\pi)^n}} \exp(-\sum_{i=1}^n \frac{(y_i - \mu_i)^2}{2\sigma^2})$$

Now, maximum likelihood estimate is simply taking as estimates those parameters which maximize the probability. In other words, the maximum-likelihood estimates for α, β_1 and β_2 are found by finding those values for those three parameters which maximize

$$P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n | \alpha, \beta_1, \beta_2),$$

where $y_1, y_2, \ldots, y_n$ are the values for the milk in our data. Now when we see our probability given above, first the expression $\frac{1}{\sigma^n \sqrt{\pi^n}}$ does not depend on the parameters α, β_1 and β_2. So, for finding which values maximize the probability we can leave that term out. Second, maximizing a function or its logarithm amounts to the same, since the logarithm is an increasing function. So, we can maximize

$$\log\Big(\sqrt{2(\pi)^n} \sigma^n \cdot P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n | \alpha, \beta_1, \beta_2) \Big) = -\sum_{i=1}^n \frac{(y_i - \mu_i)^2}{2\sigma^2}.$$

This amounts to the same as minimizing the sum or the squares

$$\sum_{i=1}^n (y_i - \mu_i)^2$$

which is precisely how out estimates where defined in the first place. So, we have proven that the maximum-likelihood estimates and our estimates $\hat{\alpha}$, $\hat{\beta}_1$ and $\hat{\beta}_2$ are identical.

Multidimensional exponential family:

We are going to show that our probability model is a multidimensional exponential family. We assume that σ is known, so it is not a parameter but just a given number. The probability model (probability distribution) for a random vector $\vec{X} = (X_1, X_2, \ldots, X_n)$ is said to be a exponential family with multidimensional parameter $\vec{\theta} = (\theta_1, \ldots, \theta_m)$ if there exists a function $\vec{T}(.)$ which maps n dimensional vectors onto m dimensional vectors

$$T(\vec{x}) = (T_1(\vec{x}), T_2(\vec{x}), \ldots, T_m(\vec{x}))$$

and such that the probability model is of the form:

$$\begin{aligned}
&P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n|\theta_1, \theta_2, \dots, \theta_m) = \\
&= c(\vec{\theta}) \cdot h(\vec{x}) \cdot \exp(\vec{T}(\vec{x}) \cdot \vec{\theta}) \; = \\
&= c(\vec{\theta}) \cdot h(\vec{x}) \cdot \exp(T_1(\vec{x})\theta_1 + \cdots + T_m(\vec{x})\theta_m),
\end{aligned}$$

where $\vec{x} = (x_1, x_2, \dots, x_n)^t$ is non-random. In the previous paragraph we have calculated the joint density function of the variables $Y_1, Y_2, \dots, Y_m$ and found the following formula:

$$P(Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n|\alpha, \beta_1, \beta_2) = \frac{1}{\sigma^n\sqrt{(2\pi)^n}} \exp(-\sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{2\sigma^2})$$

The above probability can be also written as

$$P(Y_1 = y_1, \dots, Y_n = y_n|\alpha, \beta_1, \beta_2) = \frac{1}{\sigma^n\sqrt{(2\pi)^n}} \exp\left(-\sum_{i=1}^{n}\left(\frac{y_i^2}{2\sigma^2} - \frac{2y_i \cdot \mu_i}{2\sigma^2} + \frac{\mu_i^2}{2\sigma^2}\right)\right). \tag{5.16}$$

Recalling that $\mu_i = \alpha + \beta_1 x_i^{weight} + \beta_2 x_i^{age}$, we find that the probability above in equation 5.16 is equal to:

$$\frac{1}{\sigma^n\sqrt{(2\pi)^n}} \exp\left(-\left(\frac{\sum_{i=1}^n y_i^2}{2\sigma^2} - \frac{2\sum_i^n y_i \cdot (\alpha + \beta_1 x_i^{weight} + \beta_2 x_i^{age})}{2\sigma^2} + \frac{\sum_i^n \mu_i^2}{2\sigma^2}\right)\right) = \tag{5.17}$$

$$= \frac{1}{\sigma^n\sqrt{(2\pi)^n}} \exp\left(-\frac{\sum_{i=1}^n y_i^2}{2\sigma^2} + \frac{\alpha\sum_{i=1}^n y_i + \beta_1\sum_{i=1}^n y_i x_i^{weight} + \beta_2\sum_{i=1}^n y_i x_i^{age}}{\sigma^2} - \frac{\sum_{i=1}^n \mu_i^2}{2\sigma^2}\right) = \tag{5.18}$$

$$= c(\alpha, \beta_1, \beta_2) \cdot h(y_1, \dots, y_n) \cdot \exp\left(\alpha\frac{\sum_{i=1}^n y_i}{\sigma^2} + \beta_1\frac{\sum_{i=1}^n y_i x_i^{weight}}{\sigma^2} + \beta_2\frac{\sum_{i=1}^n y_i x_i^{age}}{\sigma^2}\right) \tag{5.19}$$

where we define

$$h(y_1, \dots, y_n) := \exp(-\sum_{i=1}^{n} \frac{y_i^2}{2\sigma^2})$$

and

$$c(\alpha, \beta_1, \beta_2) := \frac{1}{\sigma^n\sqrt{(2\pi)^n}} \cdot \exp(-\frac{\sum_i^n \mu_i^2}{2\sigma^2}) = \frac{1}{\sigma^n\sqrt{(2\pi)^n}} \cdot \exp(-\frac{\sum_i^n(\alpha + \beta_1 x_i^{weight} + \beta_2 x_i^{age})^2}{2\sigma^2}).$$

When looking at expression 5.19 above, we see that this is clearly an exponential family with multidimensional parameter defined by:

$$\vec{\theta} = (\alpha, \beta_1, \beta_2)$$

and sufficient statistic given by:

$$\vec{T}(\vec{y}) = (T_1(\vec{y}), T_2(\vec{y}), T_3(\vec{y}))$$

where:

$$\begin{aligned} T_1(\vec{y}) &:= \frac{\sum_{i=1}^n y_i}{\sigma^2} = \frac{n\bar{y}}{\sigma^2} \\ T_2(\vec{y}) &:= \frac{\sum_{i=1}^n y_i x_i^{weight}}{\sigma^2} = \frac{\vec{y} \cdot \vec{x}^{weight}}{\sigma^2} \\ T_3(\vec{y}) &:= \frac{\sum_{i=1}^n y_i x_i^{age}}{\sigma^2} = \frac{\vec{y} \cdot \vec{x}^{age}}{\sigma^2} \end{aligned}$$

As usual $\vec{y}$ denotes the vector

$$\vec{y} = (y_1, y_2, \ldots, y_n)^t$$

and we used the fact, the σ here is known, so it is not considered a parameter but just a given number. Same thing for the coefficients x_i^{weight} and x_i^{age}.

The fact that we are dealing with an exponential family has an extremely important consequence: **our estimates $\hat{\alpha}$, $\hat{\beta}_1$ and $\hat{\beta}_2$ are the only possible unbiased estimates which make sense** in the current case where we consider σ known. The reason is a theorem in you book which states that for any multidimensional exponential families (in your book is simply called exponential family) the $\vec{T}(.)$ statistic is *complete* and sufficient, provided the parameters are defined on at least a rectangle. This is the case here since we have no restrictions on the parameters α, β_1 and β_2. Hence, in our current model $\vec{T}(.)$ is complete since we have seen that it is an multi-dimensional family. so, let us define the concept of completeness:

Definition 5.1 *A statistic $\vec{T}(.)$ for a probability model $P(\vec{X}|\vec{\theta}]$ with parameter $\vec{\theta}$ for a random vector $\vec{X}$ is called complete if for any two functions $f(.)$ and $g(.)$ for which:*

$$E[f(\vec{T}(\vec{X}))|\vec{\theta}] = E[g(\vec{T}(\vec{X})))|\vec{\theta}]$$

for all $\vec{\theta}$ in our parameter space, implies $f(.) = g(.)$ almost everywhere.

Let us now assume there there would be two different unbiased estimators say for α which depend only on the statistics $\vec{T}(.)$. An estimator is simply a function of the data. So, say one estimator is denoted by $f(.)$ and the other by $g(.)$. Thus, we would have $\hat{\alpha}_1 = f(\vec{T})$ and the second estimator would be $\hat{\alpha}_2 = g(\vec{T})$. Then if both estimators would be unbiased we would have

$$\alpha = E[\hat{\alpha}_1(\vec{T}(\vec{X}))|\alpha, \beta_1, \beta_2] = E[f(\vec{T}(\vec{X}))|\alpha, \beta_1, \beta_2]$$

and

$$\alpha = E[\hat{\alpha}_2(\vec{Y})|\alpha, \beta_1, \beta_2] = E[g(\vec{T}(\vec{X})|\alpha, \beta_1, \beta_2]$$

which then implies

$$E[f(\vec{T}(\vec{X}))|\alpha, \beta_1, \beta_2] = E[g(\vec{T}(\vec{X}))|\alpha, \beta_1, \beta_2] \tag{5.20}$$

for all α, β_1 and β_2. So, if the statistic $\vec{T}(.)$ in our probability model in the parameters α, β_1, β_2 is complete, equation 5.20 implies that $f(.) = g(.)$ almost everywhere. But we know by our book that any multidimensional exponential family is complete. Hence, $f(.) = g(.)$ almost everywhere and hence

$$\hat{\alpha}_1 = \hat{\alpha}_2$$

which finishes proving that there can be only one unbiased estimator for α which depends only on the statistic $\vec{T}(\vec{X})$ of the exponential family. We can prove in the same way that the unbiased estimator for β_1 and for β_2 are both unique as well. Again, we assumed σ to be known. This implies that our unbiased estimators $\hat{\alpha}$, $\hat{\beta}_1$ and $\hat{\beta}_3$ are all three unique among all unbiased estimators which depend only on $\vec{T}(\vec{X})$. since in an exponential test the statistic $\vec{T}$ is sufficient, we need only consider tests based on $\vec{T}$ for our optimal tests. So, then if we want a test to be optimal and unbiased it has to be depending only on $\vec{T}(\vec{X})$ and be unbiased and hence it is our estimates.

5.2 Estimating σ when it is not known

In our model we have that

$$\vec{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^t$$

designates a random vector with independent entries. We assume $E[\epsilon_i] = 0$ for all $i = 1, 2, \ldots, n$ and that the $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are all independent of each other. We furthermore will assume that each ϵ_i is distributed like a normal random variable $\mathcal{N}(0, \sigma)$. We are going to investigate what happens when we consider the dot product of $\vec{\epsilon}$ with a non random vectors. The dot product between a vector $\vec{a} = (a_1, \ldots, a_n)$ and a vector $\vec{b} = (b_1, b_2, \ldots, b_n)$ is defined as follows:

$$\vec{a} \cdot \vec{b} = a_1 \cdot b_1 + a_2 \cdot b_2 + \ldots + a_n \cdot b_n.$$

Furthermore, the length of a vector (Euclidean norm) is the square root of the dot product of the vector times itself:

$$|\vec{a}| = \sqrt{\vec{a} \cdot \vec{a}} = \sqrt{a_1^2 + a_2^2 + \ldots + a_n^2}.$$

Next we state and prove a lemma which gives us the main stochastic properties of the dot product of $\vec{\epsilon}$ with a non-random vector:

Lemma 5.1 *Assume the vectors $\vec{a} = (a_1, a_2, \ldots, a_n)^t$ and $\vec{b} = (b_1, b_2, \ldots, b_n)^t$ are both non random. Then we have:*

1. *First the expectation is 0:*
$$E[\vec{a} \cdot \vec{\epsilon}] = E[\vec{b} \cdot \vec{\epsilon}] = 0.$$

2. *Second the standard deviation is the length of the non-random vector up to the constant σ:*
$$\sigma_{\epsilon \cdot \vec{a}} = \sqrt{VAR[\epsilon \cdot \vec{a}]} = \sigma |\vec{a}| = \sigma \sqrt{a_1^2 + a_2^2 + \ldots + a_n^2}.$$

3. *The covariance between $\vec{\epsilon}\cdot\vec{a}$ and $\vec{\epsilon}\cdot\vec{b}$ depends only the dot product of $\vec{a}$ with $\vec{b}$:*
$$COV(\vec{a}\cdot\vec{\epsilon},\vec{b}\cdot\vec{\epsilon})=\sigma^2\vec{a}\cdot\vec{b}$$
4. *The dot product $\vec{a}\cdot\vec{\epsilon}$ has a normal distribution.*
5. *Since for normal variables, covariance equal to 0 implies independence, we get that $\vec{a}\vec{\epsilon}$ is independent of $\vec{b}\vec{\epsilon}$ if and only if $\vec{a}\cdot\vec{b}=0$, that is only when $\vec{a}$ and $\vec{b}$ are orthogonal.*

Proof. So, let us do the proofs. For the expectation we use the fact that expectation of a sum is sum of expectation:

$$\begin{aligned}
&E[\vec{a}\cdot\vec{\epsilon}]=\\
&\qquad\qquad=E[a_1\vec{\epsilon}_1+a_2\vec{\epsilon}_2+\ldots+a_n\vec{\epsilon}_n]=\\
&\qquad\qquad=E[a_1\vec{\epsilon}_1]+E[a_2\vec{\epsilon}_2]+\ldots+E[a_n\vec{\epsilon}_n]=\\
&\qquad\qquad=a_1E[\vec{\epsilon}_1+a_2E[\vec{\epsilon}_2]+\ldots+a_nE[\vec{\epsilon}_n]=a_1\cdot 0+\ldots+a_n\cdot 0=0
\end{aligned}$$

Next let us calculate the variance which is the square of the standard deviation:

$$\begin{aligned}
&\sigma^2_{\vec{\epsilon}\cdot\vec{a}}=VAR[\vec{\epsilon}\cdot\vec{a}]=\\
&=VAR[a_1\vec{\epsilon}_1+a_2\vec{\epsilon}_2+\ldots+a_n\vec{\epsilon}_n]=\\
&=VAR[a_1\vec{\epsilon}_1]+VAR[a_2\vec{\epsilon}_2]+\ldots+VAR[a_n\vec{\epsilon}_n]=\\
&=a_1^2VAR[\vec{\epsilon}_1]+a_2^2VAR[\vec{\epsilon}_2]+\ldots+a_n^2VAR[\vec{\epsilon}_n]=a_1^2\cdot\sigma^2+\ldots+a_n^2\cdot\sigma^2=\\
&\sigma^2\cdot(a_1^2+a_2^2+\ldots+a_n^2)=\sigma^2\vec{a}\cdot\vec{a}
\end{aligned}$$

next we are going to calculate the covariance of $\vec{epsilon}\cdot\vec{a}$ with $\vec{\epsilon}\cdot\vec{b}$. The covariance is bilinear. Also the covariance of independent variables is always 0. So,

$$COV(\epsilon_i,\epsilon_j)=0$$

as soon as $i\neq j$. We use this in the calculations that follow:

$$\begin{aligned}
&COV(\vec{\epsilon}\cdot\vec{a},\vec{\epsilon}\cdot\vec{b})=\\
&=COV[a_1\vec{\epsilon}_1+a_2\vec{\epsilon}_2+\ldots+a_n\vec{\epsilon}_n,b_1\vec{\epsilon}_1+b_2\vec{\epsilon}_2+\ldots+b_n\vec{\epsilon}_n]=\\
&=\sum_{i,j}a_ib_jCOV(e_i,e_j)=\sum_{i=1}^{n}a_ib_iCOV(e_i,e_i)=\sigma^2\sum_{i=1}^{n}a_ib_j=\\
&=\sigma^2\vec{a}\cdot\vec{b}
\end{aligned}$$

where we also used the fact that covariance of a variable with itself is the variance so that

$$COV(\vec{\epsilon}_i,\vec{\epsilon}_i)=VAR[\vec{\epsilon}_i]=\sigma^2.$$

The dot product $\vec{\epsilon}$ is normal simply because a linear combination of normals which are independent of each other is again normal. ∎

The next ingredient we need is the Graham-Schmidt orthogonalization theorem:

Lemma 5.2 *Assume that we have several vectors of length n given:*

$$\vec{x}^0, \vec{x}^1, \vec{x}^2, \ldots, \vec{x}^i$$

where the above vectors are all linearly independent of each other. Then there exists orthonormal vectors:

$$\vec{e}_0, \vec{e}_1, \vec{e}_2, \ldots, \vec{e}_{n-1}$$

(that is $\vec{e}_i \cdot \vec{e}_j = 0$ for all $i \neq j$ and $\vec{e}_i \cdot \vec{e}_i = 1$ for all $i = 0, 1, 2, \ldots, n-1$) so that the linear vector-subspace generated by $\vec{x}^0, \vec{x}^1, \vec{x}^2 \ldots, \vec{x}^l$ is equal to the linear subspace generated by $\vec{e}^0, \vec{e}^1, \vec{e}^2, \ldots, \vec{e}^l$ for all $l = 0, 1, 2, \ldots, i$. This is simply to say that any vector which is a linear combination of $\vec{x}^0, \vec{x}^1, \ldots, \vec{x}^l$ can also be written as a linear combination of $\vec{e}_0, \vec{e}_1, \ldots, \vec{e}_l$ for all $l \leq i$.

Proof. So, you have your sequence of factors $\vec{1}$, $\vec{x}^{weight}$, $\vec{x}^{age}$ and maybe more which you want to "orthonormalize" and the complete to find an orthonormal basis. Let us assume that we just have the three factors: $\vec{1}$, $\vec{x}^{weight}$, $\vec{x}^{age}$. We are first going to find a sequence of vectors $\vec{f}_0, \vec{f}_1, \vec{f}_2, \ldots, \vec{f}_{n-1}$ which have the properties we want except being normal (=having length 1). then we will simply take $\vec{e}_i$ to be defined as

$$\vec{e}_i = \frac{\vec{f}_i}{|\vec{f}_i|}.$$

So, how to we define the orthogonal sequence :

$$\vec{f}_0, \vec{f}_1, \vec{f}_2, \ldots, \vec{f}_{n-1}$$

The idea is very simple: we take $\vec{f}_i$ to be basically the factor number i, plus sth to make it orthogonal. The something is a linear combination of the vectors $\vec{f}_j$ with $j < i$. Here is how it works: Take $\vec{f}_0$ to be equal to the first factor. In our case:

$$\vec{f}_0 = \vec{1}.$$

Then, $\vec{f}_1$ is the factor $\vec{x}^{weight}$ minus a little sth to make it orthogonal to $\vec{f}_0$. Here is our definition:

$$\vec{f}_1 = \vec{x}^{weight} - \left(\frac{\vec{f}_0 \cdot \vec{x}^{weight}}{\vec{f}_0 \cdot \vec{f}_0} \right) \vec{f}_0$$

We can verify that $\vec{f}_1$ is orthogonal to $\vec{f}_0$:

$$\begin{aligned} &\vec{f}_0 \cdot \vec{f}_1 = \\ &= \vec{f}_0 \cdot (\vec{x}^{weight} - \left(\frac{\vec{f}_0 \cdot \vec{x}^{weight}}{\vec{f}_0 \cdot \vec{f}_0} \right) \vec{f}_0) = \\ &= \vec{f}_0 \cdot \vec{x}^{weight} - \left(\frac{\vec{f}_0 \cdot \vec{x}^{weight}}{\vec{f}_0 \cdot \vec{f}_0} \right) (\vec{f}_0 \cdot \vec{f}_0) = 0 \end{aligned}$$

so the fectors $\vec{f_0}$ and $\vec{f_1}$ are orthogonal. Note also, that any vector which can be expressed as a combination of $\vec{1}$ and $\vec{x}^{weight}$ can be written as a linear combination of $\vec{f_0}$ and $\vec{f_1}$:

$$\begin{aligned}
&a\vec{1} + b\vec{x}^{weight} = \\
&= a\vec{f_0} + b(\vec{f_1} + \left(\frac{\vec{f_0} \cdot \vec{x}^{weight}}{\vec{f_0} \cdot \vec{f_0}}\right)\vec{f_0}) = \\
&(a + b\frac{\vec{f_0} \cdot \vec{x}^{weight}}{\vec{f_0} \cdot \vec{f_0}}) \cdot \vec{f_0} + b\vec{f_1}
\end{aligned}$$

Now, to define $\vec{f_2}$ we simply take the third factor $\vec{x}^{age}$ and add a combination of $\vec{f_0}$ and $\vec{f_1}$ to make it orthogonal:

$$\vec{f_2} := \vec{x}^{age} - \vec{f_0}\frac{\vec{x}^{age} \cdot \vec{f_0}}{\vec{f_0} \cdot \vec{f_0}} - \vec{f_1}\frac{\vec{x}^{age} \cdot \vec{f_1}}{\vec{f_1} \cdot \vec{f_1}}.$$

One can now check that $\vec{f_2}$ is orthogonal to $\vec{f_1}$ and $\vec{f_0}$:

$$\begin{aligned}
&\vec{f_1} \cdot \vec{f_2} := \\
&= \vec{f_1} \cdot (\vec{x}^{age} - \vec{f_0}\frac{\vec{x}^{age} \cdot \vec{f_0}}{\vec{f_0} \cdot \vec{f_0}} - \vec{f_1}\frac{\vec{x}^{age} \cdot \vec{f_1}}{\vec{f_1} \cdot \vec{f_1}}) \\
&= \vec{f_1} \cdot \vec{x}^{age} - (\vec{f_1} \cdot \vec{f_0})\frac{\vec{x}^{age} \cdot \vec{f_0}}{\vec{f_0} \cdot \vec{f_0}} - \vec{f_1} \cdot \vec{f_1}\frac{\vec{x}^{age} \cdot \vec{f_1}}{\vec{f_1} \cdot \vec{f_1}} = 0.
\end{aligned}$$

Similary we can show that $\vec{f_2}$ and $\vec{f_0}$ are also orthogonal. Once, we are done with the factors, we take any vector which is not a combination of the already defined $\vec{f_0}, \vec{f_1}, \vec{f_2}$ call that vector $\vec{z}$. Then define $\vec{f_3}$ to be

$$\vec{f_3} = \vec{z} - \vec{f_0}\frac{z \cdot \vec{f_0}}{\vec{f_0} \cdot \vec{f_0}} - \vec{f_1}\frac{z \cdot \vec{f_1}}{\vec{f_1} \cdot \vec{f_1}} - \vec{f_2}\frac{z \cdot \vec{f_2}}{\vec{f_2} \cdot \vec{f_1}}$$

So, then once we have $\vec{f_3}$ we go on ■

We are now ready to apply the above lemma to find an unbiased estimator for σ: so let us consider the factor vectors $\vec{\mathbf{1}}$, $\vec{x}^{weigth}$ and $\vec{x}^{age}$ for our wonderful swiss cows. Let now $\vec{e}_1, \vec{e}_2, \ldots, \vec{e}_n$ be the sequence (basis) of orthonormal vectors which the above lemma 5.2 garanties the existence of. For this we take

$$\vec{x}^0 = \vec{\mathbf{1}}, \vec{x}^1 = \vec{x}^{weight}, \vec{x}^2 := \vec{x}^{age}$$

and $j = 2$. So, then we have $\vec{e}_1, \vec{e}_2, \ldots$ are orthonormal and $\vec{e}_0$ and $\vec{\mathbf{1}}$ are colinear. Furthermore, any vector which can be wrtten as $a\vec{1} + b\vec{x}^{weight}$ can be written as a linear combination of $\vec{e}_0$ and $\vec{e}_1$. Finally, any vector which can be written as a

$$a\vec{1} + b\vec{x}^{weight} + c\vec{x}^{age}$$

can be written as a linear combination of $\vec{e}_0, \vec{e}_1, \vec{e}_2$. This is the same as saying that $\vec{e}_0$, $\vec{e}_1$ and $\vec{e_2}$ generated the same subspace of vectors as the three vectors: $\vec{\mathbf{1}}$, $\vec{x}^{weight}$ and $\vec{x}^{age}$. Also, all the vectors $\vec{e}_i$ are orthogonal to each other. Let us project our vector $\vec{y}$ onto $\vec{e^j}$ with $j \geq 3$. Then we have in the current example that all the factor vectors $\vec{\mathbf{1}}$, $\vec{x}^{weight}$ and $\vec{x}^{age}$ are orthogonal to $\vec{e^j}$ for $j \geq 3$. We can use this orthogonality to find: the dot product $\vec{y} \cdot \vec{e}_j$ is equal to

$$\begin{aligned}\vec{y} \cdot \vec{e}_j &= \left(\alpha\vec{\mathbf{1}} + \beta_1\vec{x}^{weight} + \beta_2\vec{x}^{age} + \vec{\epsilon}\right) \cdot \vec{e}_j = \\ &= \alpha\vec{\mathbf{1}} \cdot \vec{e}_j + \beta_1\vec{x}^{age} \cdot \vec{e}_j + \beta_2\vec{x}^{age} \cdot \vec{e}_j + \vec{\epsilon} \cdot \vec{e}_j = 0 + \vec{\epsilon} \cdot \vec{e}_j = \vec{\epsilon} \cdot \vec{e}_j.\end{aligned}$$

But, now we have shown that

$$\vec{y} \cdot \vec{e}_j = \vec{\epsilon} \cdot \vec{e}_j. \tag{5.21}$$

So, this implies first of all that the coefficients $\vec{y} \cdot \vec{e}_j$ do not depend on any parameter α β_1 or β_2 nor on the factor vector, since $\vec{\epsilon} \cdot \vec{e}_j$ does not. Second, the dot product on the right side of 5.21 is exactly of the form our lemma 5.2 is about. So we can apply the results of lemma 5.2 and get: Since the vectors $\vec{e}_j$ are orthogonal among each other, we get that the coefficients $\vec{\epsilon} \cdot \vec{e}_j = \vec{y} \cdot \vec{e}_j$ are independent of each other. Furthermore, because the vectors $\vec{e^j}$ have length one, the standard deviation of $\vec{\epsilon} \cdot \vec{e^j}$ is σ. The expectation is 0, so that

$$E[\vec{\epsilon} \cdot \vec{e^j}] = 0,$$

for $j \geq 3$.
Of course the coefficient are normal. Hence, when we consider the random vector

$$(\vec{y} \cdot \vec{e}_3, \vec{y} \cdot \vec{e}_4, \vec{y} \cdot \vec{e}_4, \ldots, \vec{y} \cdot \vec{e}_n),$$

(which is the projection of $\vec{y}$ onto the space orthogonal to the factor space), we get a vector with i.i.d entries which are all normal $\mathcal{N}(0, \sigma)$. For such a vector the best unbiased estimate of the variance σ^2 is given by taking the sum of the squares:

$$\hat{\sigma}^2 := \frac{\vec{y} \cdot \vec{e}_2^2 + \vec{y} \cdot \vec{e}_3^2 + \vec{y} \cdot \vec{e}_4^2 + \ldots + (\vec{y} \cdot \vec{e}_{n-1}^2}{n-3}.$$

In general with k-factors instead of just 3 we would get the estimator

$$\hat{\sigma}^2 := \frac{(\vec{y} \cdot \vec{e}_{k-1})^2 + (\vec{y} \cdot \vec{e}_k)^2 + (\vec{y} \cdot \vec{e}_{k+1})^2 + \ldots + (\vec{y} \cdot \vec{e}_{n-1})^2}{n-k}.$$

The estimate for the standard deviation is then

$$\hat{\sigma} := \sqrt{\frac{(\vec{y} \cdot \vec{e}_{k-1})^2 + (\vec{y} \cdot \vec{e}_k)^2 + (\vec{y} \cdot \vec{e}_{k+1})^2 + \ldots + (\vec{y} \cdot \vec{e}_{n-1})^2}{n-k}}.$$

Why do we say best "unbiased estimator"? Here is the answer: let us introduce the notation $N_i := \vec{y} \cdot \vec{e}_i$ for $i = 0, 1, 2, \ldots, n-1$. Then we have that the coefficients N_i for

$i \geq k-1$, (when we have k factors) all have expectation 0. In that case, the variance is simply the expectation of the square:

$$\sigma^2 = VAR[N_i] = E[(N_i)^2]$$

for $i \geq k$. So estimating the variance of N_i is the same as estimating the expectation of N_i^2. But for estimating the expectation of a bunch of variables the average is always an unbiased estimator since

$$E[\frac{Z_1 + Z_2 + \ldots + Z_m}{m}] = \frac{1}{m} E[Z_1 + \ldots + Z_m] =$$
$$= \frac{1}{m}(E[Z_1] + E[Z_2] + \ldots + E[Z_m]) = \frac{mE[Z_1]}{m} = E[Z_1]$$

where we assumed that all the variables Z_i have same expectation. In our setting $Z_i := N_i^2$ we find that

$$\hat{\sigma}^2 = \frac{N_k^2 + N_{k+1}^2 + \ldots + N_{n-1}^2}{n-k}$$

is an unbiased estimator of the expectation

$$E[Z_i] = E[N_i^2] = VAR[N_i] = \sigma^2.$$

Now the joint density of $N_k, N_k, N_{k+1}, \ldots, N_{n-1}$ is given by

$$\frac{1}{(\sqrt{2\pi\sigma^2})^{n-k}} \exp\left(-\sum_{i=k}^{n-1} N_i^2 \cdot \frac{1}{2\sigma^2}\right).$$

So, we are dealing with an exponential family with parameter $-\frac{1}{2\sigma^2}$ and statistic $T(.)$ equal to:

$$T(N_k, N_k, N_{k+1}, \ldots, N_{n-1}) := N_k^2 + \ldots + N_{n-1}^2$$

Indeed the joint density can be written as:

$$c(\frac{-1}{2\sigma^2}) \exp\left(T(N_{k-1}, N_k, N_{k+1}, \ldots, N_n) \cdot \frac{-1}{2\sigma^2}\right),$$

where $c(x)$ is the function

$$c(x) = \left(\sqrt{-\frac{x}{\pi}}\right)^{n-k}.$$

For an exponential family, the statistic $T(.)$ is sufficient, so we can through out everthing else. Hence, we can base any optimal estimator solely on the statistic $T(N_k, N_k, N_{k+1}, \ldots, N_n)$. But for an exponential family, there is only one unbiased estimator based on $T(.)$, because of completeness as discussed earlier. But, our estimator is unbiased and only based on $T(N_k, N_k, N_{k+1}, \ldots, N_{n-1}))$. So in that sense, it is the best unbiased estimator possible. Note that by definition we have

$$\hat{\alpha}\vec{1} + \hat{\beta}_1 \vec{x}^{weigth} + \hat{\beta}_2 \vec{x}^{age}$$

Now, note that we have seen that the factor vectors $\vec{1}$, $\vec{x}^{weigth}$ and $\vec{x}^{age}$ are perpendicular to

$$\vec{y} - (\hat{\alpha} + \hat{\beta}_1\vec{x}^{weight} + \hat{\beta}_2\vec{x}^{age})$$

This is given by the equations 5.9, 5.10 and 5.11. The explanation of why this holds is simple: the estimates $\hat{\alpha}$, $\hat{\beta}_1$ and $\hat{\beta}_2$ Are simply this values for α, β_1 and β_2 which minimise the sum of the error squares:

$$\sum_{i=1}^{n}(y_i - \alpha - \beta_1 x_i^{weight} - \beta_2 x_i^{age})^2$$

This is the same as finding α, β_1 and β_2 which minimize the distance square between the vector $\vec{y}$ and

$$\alpha - \beta_1\vec{x}_i^{weight} - \beta_2\vec{x}^{age}. \tag{5.22}$$

But when we consider all points which can be written like 5.22 (that is for which there is any value of α, β_1 and β_2 so that expression 5.22 equals the point) then this collectionof point is a linear subvector space. In three dimensions for example, all vector subspaces are either $\{\vec{0}\}$, or a line going through the origin or a plane going through the origin or the whole space. But if I wan to go in shortest path from a point in space $\vec{y}$ to a plane or a line, the shortest path is always perpendiculr to the line or to the plane to which I want to go. Same things holds in higher dimensional vector space: shortest path from a point $\vec{y}$ to a subspace is by going from $\vec{y}$ to the subspace in orthogonal way to the subspace. Now for orthogonal vectors we can use the good old pythagoras: So we have

$$\vec{y} = \left[\vec{y} - (\hat{\alpha} + \hat{\beta}_1\vec{x}^{weigth} + \hat{\beta}_2\vec{x}^{age})\right] + \left(\alpha - \beta_1 x_i^{weight} - \beta_2 x_i^{age}\right)$$

Note that the first part of the sum on the right side of the above equation is equal to

5.3 Another estimate for σ: maximum-Lieklyhood.

Unlike many other situations, here there can be a big difference between maximum-lieklyhood estimate for σ an the best unbiased estimate. Let us see what the maximum-Likelyhood estimate for σ is: the joint density function for $Y_1, Y_2, \ldots, Y_n$ is obtained as mentionned already by mutiplying the different densities with each other since the coefficients Y_i are independent of each other. Now Y_i is normal with expectation

$$\mu_i; = \alpha + \beta_1 x_i^{weight} + \beta_2 x_i^{age}$$

so the density of Y_i at the point y_i is given by

$$\frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{y_i - \mu_i)^2}{2\sigma^2}\right)$$

so the joint density which is the product of the density of each Y_i is given by

$$f_{Y_1,Y_2,\ldots,Y_n}(y_1, y_2, \ldots, y_n) = \frac{1}{\sqrt{(2\pi)^n}\sigma^n}\exp\left(-\sum_{i=1}^{n}\frac{(y_i - \mu_i)^2}{2\sigma^2}\right)$$

we want to find σ which maximizes the density above for given numeric values $y_1, y_2, \ldots, y_n$. This is the same as maximizing the logarithm of the density:

$$-n\ln(\sigma) - (\sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{2\sigma^2}),$$

which is the same as minimizing

$$+n\ln(\sigma) + (\sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{2\sigma^2}) = n\ln(\sigma) + \sum_{i=1}^{n} \frac{(y_i - \alpha - \beta_1 x_i^{weigth} - \beta_2 x_i^{age})^2}{2\sigma}$$

Now the above density also depends on the parameter α, β_1 and β_2. But we have seen, that the density gets maximized for given fixed value of σ with the estimates $\hat{\alpha}$, $\hat{\beta}_1$ and $\hat{\beta}_2$. So, to find the value of σ maximizing the likelyhood we can replace in the formula for density α, β_1, β_2 by $\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2$. So, hence, we want to maximize

$$n\ln(\sigma) + (\sum_{i=1}^{n} \frac{(y_i - \hat{\alpha} - \hat{\beta}_1 x_i^{weigth} - \hat{\beta}_2 x_i^{age})^2}{2\sigma^2}$$

to minimize the expression above we can now take the derivative according to σ and set it equal to 0:

$$\frac{n}{\sigma} - (\sum_{i=1}^{n} \frac{(y_i - \hat{\alpha} - \hat{\beta}_1 x_i^{weigth} - \hat{\beta}_2 x_i^{age})^2}{\sigma^3} = 0$$

which is equivalent to

$$\sigma^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\alpha} - \hat{\beta}_1 x_i^{weigth} - \hat{\beta}_2 x_i^{age})^2}{n}$$

So, the maximum-lieklihood estimate is given by

$$\mathcal{MLE}(\sigma^2) := \sum_{i=1}^{n} \frac{(y_i - \hat{\alpha} - \hat{\beta}_1 x_i^{weigth} - \hat{\beta}_2 x_i^{age})^2}{n}$$

Note that this estimate is very similar to the unbaised estimate $\hat{\sigma}$: the only difference is that here we devide by n instead of n minus the number of factors! So, if the number of factors will be small the two estimates are very similar. But if the number of factors is close to the dimension of the space, then the two might be somewhant different.

5.4 Precision of the estimates of the parameters

How precise are our estimates? We have to figure out the standard deviations. When the factor vector are strongly correlated we find that our estimates can be very imprecise. Most of the time we will take the values of the factor minus the data average: so for example for the factor $\vec{x}^{age}$ we do not take as i-th entry the value x_i^{age} which is the age of the i-th cow. Instead, we take $x_i^{age} - \bar{x}^{age}$ where the age average is defined by:

$$\bar{x}^{age} := \sum_{i=1}^{n} x_i^{age}.$$

So taking the vector $\vec{x}^{age}$ to be equal to

$$\vec{x}^{age} = (x_1^{age} - \bar{x}^{age}, x_2^{age} - \bar{x}^{age}, \ldots, x_n^{age} - \bar{x}^{age})^t$$

we get that it is perpendicular to $\vec{1}$. Furthermore, in that case the length of the vector is equal to the *sqrtn* times the sample standard deviation of the factor "age", because

$$|\vec{x}^{age}| = \sqrt{\vec{x}^{age} \cdot \vec{x}^{age}} = \sqrt{n} \cdot \sqrt{\sum_{i=1}^{n} (x_i^{age} - \bar{x}^{age})^2 / n} = \sqrt{n} \cdot sd(\texttt{age}).$$

Let us summarize how precise the estimate $\hat{\beta}^{age}$ is depending on wether the factors are orthogonal or not:

1. Assume first that the factor $\vec{\mathbf{1}}$, $\vec{x}^{weigth}$ and $\vec{x}^{age}$ are orthogonal. Then

   $$\sigma_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{n}|\vec{x}^{age}|} = \frac{\sigma}{\sqrt{n}\sqrt{\vec{x}^{age} \cdot \vec{x}^{age}}} = \frac{\sigma}{\sqrt{n}\texttt{sd(age)}}.$$

   so the precsion behave like $1/\sqrt{n}$ times constant.

2. If the factor vector are close to each other, then the estimates precision could be very bad. Assume for example the the distance of the rescaled vector $\frac{\vec{x}^{age}}{|\vec{x}^{age}|}$ to the linear subspace generated by the other factors is $\delta > 0$. (In the present case, the distance to the space generated by $\vec{1}$ and $\vec{x}^{age}$). Then the standard deviation is given by

   $$\sigma_{\hat{\beta}_1} = \frac{\sigma}{\delta\sqrt{n} \cdot \texttt{sd(age)}} \tag{5.23}$$

Let us give the proof. So, if the distance of $\vec{x}^{age}$ to the linear subspace generated by the other factors is $\delta \cdot |\vec{x}^{age}|$ and considering again our orthonormal sequence $\vec{e}_0, \vec{e}_1, \ldots$ we get that $\vec{x}^{age}$ can be written as $\delta|\vec{x}^{age}|\vec{e}_2$ plus a linear combination of the other factors. So,

$$\vec{x}^{age} = \delta|\vec{x}^{age}| \cdot \vec{e}_2 + a\vec{1} + b\vec{x}^{weight} \tag{5.24}$$

for some real coefficients a and b. From the three equations 5.3, 5.10 and 5.11 it follows that any linear combination $a\vec{1} + b\vec{x}^{weight} + c\vec{x}^{age}$ is yorthogonal to

$$\vec{y} - \hat{\alpha}\vec{1} - \hat{\beta}_1\vec{x}^{weigth} - \hat{\beta}_2\vec{x}^{age}. \tag{5.25}$$

But clearly $\vec{e}_2$ can be written as linear combination of the three factors $\vec{1}$, $\vec{x}^{weight}$ and $\vec{x}^{age}$. Hence $\vec{e}_2$ must be orthogonal to 5.25, so that

$$\vec{e}_2 \cdot \left(\vec{y} - \hat{\alpha}\vec{1} - \hat{\beta}_1\vec{x}^{weigth} - \hat{\beta}_2\vec{x}^{age}\right) = \vec{0}$$

we can now replace in the equation above $\vec{y}$ by $\alpha\vec{1} + \beta_1\vec{x}^{weight} + \beta_2\vec{x}^{age} + \vec{\epsilon}$ to obtain:

$$\vec{e}_2 \cdot \left((\alpha - \hat{\alpha})\vec{1} + (\beta_1 - \hat{\beta}_1)\vec{x}^{weigth} + (\beta_2 - \hat{\beta}_2)\vec{x}^{age} + \vec{\epsilon}\right) = \vec{0} \qquad (5.26)$$

So by $\vec{e}_2$ being orthogonal to the factors $\vec{1}$ and $\vec{x}^{weight}$ equation 5.26 finally becomes

$$(\beta_2 - \hat{\beta}_2)\vec{e}_2 \cdot \vec{x}^{age} = -\vec{e}_2 \cdot \vec{\epsilon}$$

so that our estimation error for β_2 can be written as

$$\beta_2 - \hat{\beta}_2 = -\frac{\vec{e}_2 \cdot \vec{\epsilon}}{\vec{e}_2 \cdot \vec{x}^{age}} \qquad (5.27)$$

by equation 5.24 we get that

$$\vec{e}_2 \cdot \vec{x}^{weight} = \delta \cdot |\vec{x}^{weight}|$$

and hence equation 5.27 becomes

$$\beta_2 - \hat{\beta}_2 = -\frac{\vec{e}_2 \cdot \vec{\epsilon}}{\delta\sqrt{n} \cdot \texttt{sd(age)}}$$

Now, the standard deviation of a variable times a constant is equal to the constant times the standard deviation. This implies

$$\sigma_{\hat{\beta}_1} = \sigma_{-\frac{\vec{e}_2\vec{\epsilon}}{\delta\sqrt{n}sd(age)}} = \frac{\sigma_{\vec{e}_2 \cdot \vec{\epsilon}}}{\delta\sqrt{n} \cdot \texttt{sd(age)}} = \frac{\sigma}{\delta\sqrt{n} \cdot \texttt{sd(age)}},$$

where we used lemma 5.21 for obtaining the last equation above.
Now, we can find a confidence interval for β_2 using the above standard deviation: note that $\hat{\beta}_2$ is a normal variable with expectation β_1. But for a normal variable we know how to determine the confidence interval. For example, the 95% symmetric confidence interval for β_1 is then given by

$$[\hat{\beta}_1 - 1.96\sigma_{\hat{\beta}_1}, \hat{\beta}_1 + 1.96\sigma_{\hat{\beta}_1}] = [\hat{\beta}_1 - 1.96\frac{\sigma}{\delta\sqrt{n} \cdot \texttt{sd(age)}}, \hat{\beta}_1 + 1.96\frac{\sigma}{\delta\sqrt{n} \cdot \texttt{sd(age)}}] \qquad (5.28)$$

Now, in the case that you don't know σ, you will simply replace it by the estimate $\hat{\sigma}$ which we defined earlier. Then becaue of these additional source of "error" (not having the ture σ bu only an estimate), your confidence interval 5.28 will be sligthly too small. so, instead of the coefficient 1.96 form the normal table you take a slightly bigger coefficient which is to be found in a student t-table with $n - k$ degrees of freedom. The confidence interval is then given by:

$$[\hat{\beta}_1 - t_{0.05}\frac{\hat{\sigma}}{\delta\sqrt{n} \cdot \texttt{sd(age)}}, \hat{\beta}_1 + t_{0.05}\frac{\hat{\sigma}}{\delta\sqrt{n} \cdot \texttt{sd(age)}}]$$

5.4.1 Testing for $\beta_2 = 0$

Again, $\hat{\beta}_2$ is a normal variable with expectation β_1. We know how to test for $\beta_1 = 0$ on a given significance level for such a variable (assuming fist σ to be known). The two-sided test with significane $\alpha > 0$ used to test $\mu = 0$ agains $\mu \neq 0$ is equivalent to calculating the confidence interval on the confidence level $1 - \alpha$ and then rejecting the hypothesis $\beta_1 = 0$ if 0 is not in the confidence interval. So, calculating the symmetric confidence interval and then rejecting the hypothesis if 0 is not in the confidence interval has an advantage: when you fail to reject the hypothesis $\beta_2 = 0$ you might wonder why it is. Maybe it is because you don't have enough precision, that is there is too big standard deviation to the estimate. Or maybe it is becauser really β_0 is close to 0. Imagine for example that some cows eat powerfood and you want to see if this improves milk production. So you need to know if the coefficient $\beta_{powerfood}$ is 0 or not. Say you fail to reject $\beta_{powerfood} = 0$. Then you look at the confidence interval and you will see: if it is very big, then the problem is lack of precision in the estimate most probability due to colinearity of the factors, that is small δ making the standard deviation big. So, then you can try to increase the number of cows to get the $\sqrt{n}$ in the standard deviation better. this might be very hard work: to get the precision 10 times better you will need 100 times more cows!

5.5 Predicting the amount of milk of a new cow

there two type of problems:

1. Finding out if a factor affect the milk production. for example is age important for how much milk a cow produces. In other words, this is about testing the hypothesis $\beta_2 = 0$ agains $\beta_2 \neq= 0$.

2. The second problem is predicting. Assume that you buy a new cow online and they tell you the weigth and the age but not how much milk she produces. Then you want to estimate how much milk this cow is going to produce and give a confidence interval.

Say you are a big specialist with cows: you sell them and buy them for a living. Then it could be that you know the true parameters α, β_1 and β_2 as well as σ. Now the amount of milk prodced by the new cow is

$$Y_{new} = \alpha + \beta_1 x_{new}^{weight} + \beta_2 x_{new}^{age} + \epsilon_{new}$$

where as usual $E[\epsilon_{new}] = 0$. so, the expected abount of milk is

$$\begin{aligned} E[Y_{new}] =& E[\alpha + \beta_1 x_{new}^{weight} + \beta_2 x_{new}^{age} + \epsilon_{new}] = \\ =& E[\alpha] + E[\beta_1 x_{new}^{weight}] + E[\beta_2 x_{new}^{age}] + E\epsilon_{new}] = \qquad \alpha + \beta_1 x_{new}^{weight}] + \beta_2 x_{new}^{age} \end{aligned}$$

where we used that the expectation of a constant is the constant itself. So, you know the expected amount of milk. the standard deviation is σ. So, in the case that you know all the parameters because you are a big experts who has been working with cows for years, your 95% confidence interval is going to be

$$[\alpha + \beta_1 x_{new}^{weight}] + \beta_2 x_{new}^{age} - 1.96 \cdot \sigma, \alpha + \beta_1 x_{new}^{weight}] + \beta_2 x_{new}^{age} + 1.96 \cdot \sigma].$$

If you don't know the parameters α, β_1 and β_2 exactly you will have to use the estimates instead. This is then adding some uncertainty. So you replace

$$\alpha + \beta_1 x_{new}^{weight}] + \beta_2 x_{new}^{age}$$

by

$$\hat{\alpha} + \hat{\beta}_1 x_{new}^{weight}] + \hat{\beta}_2 x_{new}^{age}$$

this adds an estimation error equal to

$$error_{new} := \alpha - \hat{\alpha} + \beta_1 - \hat{\beta}_1 x_{new}^{weight}] + (\beta_2 - \hat{\beta}_2) x_{new}^{age}$$

As estimate for the amount of milk produced by the new cow we use

$$\hat{Y}_{new} := \hat{\alpha} + \hat{\beta}_1 x_{new}^{weight}] + \hat{\beta}_2 x_{new}^{age}$$

This is an unbiased estimator for the expected amount of milk for the new cow given her size and age. This is so, since the estimators of the parameters are unbiased:

$$\begin{aligned} E[\hat{Y}_{new}] = & \qquad\qquad\qquad = E[\hat{\alpha} + \hat{\beta}_1 x_{new}^{weight} + \hat{\beta}_2 x_{new}^{age}] = \\ & = E[\hat{\alpha}] + E[\hat{\beta}_1] x_{new}^{weight} + E[\hat{\beta}_2] x_{new}^{age}] \\ & = \alpha + \beta_1 x_{new}^{weight} + \beta_2 x_{new}^{age} \end{aligned}$$

the confidence interval for the milk which the new cow is coing to produce could be view roughly as the estimate $\hat{Y}_{new}$ plus minus two times the standard deviation of the estimation error. More precisely the 95% confidence interval is going to be

$$[\hat{Y}_{new} - 1.96\sqrt{\sigma^2_{\hat{Y}_{new}} + \sigma^2}, \hat{Y}_{new} + 1.96\sqrt{\sigma^2_{\hat{Y}_{new}} + \sigma^2}]$$

So, we need to determine

$$\sigma_{\hat{Y}_{new}}$$

which can be viewed as the typical error if we want to estimate the expected amoung of milk of the new cow rather than the milk it will actually produce. the imprecision could be very bad. But let us first look at a cow which is typcial: average age and average weight in our data set. We use equation 5.6 and find:

$$\sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta}_1 x_i^{weigth} - \hat{\beta}_2 x_i^{age}) = 0 \tag{5.29}$$

$$\frac{\sum_{i=1}^{n} y_i}{n} - \hat{\alpha} - \hat{\beta}_1 \frac{\sum_{i=1}^{n} x_i^{weigth}}{n} - \hat{\beta}_2 \frac{\sum_{i=1}^{n} x_i^{age}}{n}) = 0 \tag{5.30}$$

$$\bar{y} - \hat{\alpha} - \hat{\beta}_1 \bar{x}^{weight} - hat\beta_2 \bar{x}^{age} = 0 \tag{5.31}$$

Where $\bar{y}$ designates the average of the y_i's in our data and bar is always used to designate a sample average. Since

$$y_i = \alpha + \beta_1 x_i^{weigth} + \beta_2 x_i^{age} - \epsilon_i$$

we can sum these equation over $i = 1$ up to $i = n$ and then divide by n. We then find that

$$\bar{y} = \alpha + \beta_1 \bar{x}^{weight} + \beta_2 \bar{x}^{age} + \frac{\sum_{i=1}^n \epsilon_i}{n}$$

Combining the last equation above with **??**, we find finally that the difference

$$\alpha - \hat{\alpha} + (\beta_1 - \hat{\beta}_1)\bar{x}^{weight} + (\beta_2 - \hat{\beta}_2)\bar{x}^{age} = -\frac{\sum_i^n \epsilon_i}{n} \tag{5.32}$$

Note that the right side of the equation above is the difference between the expected milk and the estimate of what the expected milk should be for a new cow with average values!!! That is if $x_{new}^{weight} = \bar{x}^{new}$ and if $x_{new}^{age} = \bar{x}^{age}$, then the expected milke for the new cow will be

$$\alpha + \beta_1 x_{new}^{weight} + \beta_2 x_{new}^{age} = \alpha + \beta_1 \bar{x}^{weight} + \beta_2 \bar{x}^{age}.$$

The estimate for the expected milk of the new cow (not the estimate for the actual milk, just for the expected milk given size and age of the new cow) is

$$\hat{\alpha} + \hat{\beta}_1 x_{new}^{weight} + \hat{\beta}_2 x_{new}^{age} = \hat{\alpha} + \hat{\beta}_1 \bar{x}^{weight} + \hat{\beta}_2 \bar{x}^{age}.$$

So, the estiamtion error for the expected milk for the new cow is the difference between the true expected milk and the estimate is given by

$$\alpha + \beta_1 \bar{x}^{weight} + \beta_2 \bar{x}^{age} - (\hat{\alpha} + \hat{\beta}_1 \bar{x}^{weight} + \hat{\beta}_2 \bar{x}^{age}) =$$
$$= \alpha - \hat{\alpha} + (\beta_1 - \hat{\beta}_1)\bar{x}^{weight} + (\beta_2 - \hat{\beta}_2)\bar{x}^{age} = -\frac{\sum_i^n \epsilon_i}{n}$$

where we used **??** Now for the average of the ϵ_i the standard deviation is simply the standard deviation of one of the terms divied by $\sqrt{n}$

$$VAR[\frac{\sum_i^n \epsilon_i}{n}] = \frac{\sum_i^n VAR[\epsilon_i]}{n^2} = \frac{n \cdot VAR[\epsilon_1]}{n^2} = \frac{\sigma^2}{n}$$

and hence taking the square root on both sides of the above equation and using that the standard deviation is the square root of the variance we find

$$\sigma_{(\epsilon_1 + \ldots + \epsilon_n)/n} = \frac{\sigma}{\sqrt{n}}.$$

So, we know the precision of the estimate fo the new cows expected milk production if the cow is very typical. If it is not then the calculation is more complicated and we will do it later. So, we have that

$$\sigma_{\hat{Y}_{new}} = \frac{\sigma^2}{\sqrt{n}}$$

when the new cow has age and height exactly equal to the average values in our data set. So, we can use this value now for giving the confidence interval for such a super typical cow!

5.6 Testing for several coefficients β_i to be zero: ANOVA

Say we work with weight, age, but also for example size. Maybe both age and size are superfluous for predicting milk production of a cow. So, we consider the model:

$$milk = \alpha + \beta_1 \cdot weight + \beta_2 \cdot age + \beta_3 \cdot size + cow.nb.i.error$$

or

$$y_i = \alpha + \beta_1 \cdot x_i^{weight} + \beta_2 \cdot x_i^{age} + \beta_3 x_i^{size} + \epsilon_i,$$

where as usual y_i is the mount of milk produced by the i-th cow. As usual x_i^{weight}, x_i^{age} resp. x_i^{size} denote the weight, age, resp. size of the i-th cow. Then ϵ_i is as always the individual fluctuation term of the i-th cow. In vector format, we have thus

$$\vec{y} = \alpha\vec{1} + \beta_1 \vec{x}^{weight} + \beta_2 \vec{x}^{age} + \beta_3 \vec{x}^{size} + \vec{\epsilon}$$

We want to test the hypothesis

$$H_0 : \beta_2 = \beta_3 = 0.$$

If this hypothesis where true, it would mean, that in order to predict milk production in a cow, once we use the weight factors, we can drop size and age.

Another such multi-factor situation, could be when try to test if some power-food has some effect on milk production. Assume that in our data-set with cows, we have that to some cows a power food I is given and to some others a second type of power food. Some cows don't receive any power food. This would then lead to two factor vectors: $\vec{x}^{powerfoodI}$ and $\vec{x}^{powerfoodII}$. the i-th entry of $\vec{x}^{powerfoodI}$, that is $x_i^{powerfoodI}$ would then be equal to 1 if the i-th cow has received power food number I and would be 0 otherwise. We would define the vector $\vec{x}^{powerfoodII}$ in a similar manner, but using the second power food instead of the first. So, our model would be

$$\vec{y} = \alpha\vec{1} + \beta_1 \vec{x}^{weight} + \beta_2 \vec{x}^{age} + \beta_3 \vec{x}^{size} + \beta_4 \vec{x}^{powerfoodI} + \beta_5 \vec{x}^{powerfoodII} + \vec{\epsilon}.$$

in that model we may want to test for the hypothesis $H_0 : \beta_3 = \beta_4 = 0$. If we can reject the hypothesis, on a certain significance level, than that would mean that we have proven that the power foods have some effect (on the given significance level at which we are testing). More precisely, we would have proven that at least one of them has a significant effect on milk production. Why do we also use the other factors: weight, age and size? The reason is that this allows us to reduce the variance in the problem: it is very possible that the difference of weight between the cows, creates a enormous disparity in the milk production from one cow to the other. Maybe the power food has some effect on milk production, but less than the weight of the cow. In that case, we could be in a situation where the effect of the power food is not seen right away, because it is masked by the big fluctuation of the amount of milk produced from one cow to the next due to the different weights of the cows. But, now, when we include the weight of the cow into the model, this can lead to "take away" the big fluctuation in milk which is due to size, and then leave us with a smaller fluctuation (smaller average residuals) once the contribution of size is taken out. And in that case, it is quite likely, that if the power food has indeed an effect on milk production (and it is not a hoax), then we will be able to detect it

So let us go back to our model with only weight, age and size. Now, we have seen how to test for one coefficient to be equal to 0. Testing for several coefficients to be 0 at the same time is not exactly the same thing. Specially if there are many factors involved in the test. The reason is as follows: assume that we would try to test for 20 coefficients to be all 0, so that H_0 would be equal to:

$$\beta_2 = \beta_3 = \beta_4 = \ldots = \beta_{22} = 0.$$

Then if we have a critical value $c_i^{0.05}$ at the 5% significance level for each of these coefficients. So,

$$P(|\hat{\beta}_i| \geq c_i^{0.05} | \beta_i = 0) = 0.05$$

for each $i = 2, 3, \ldots, 22$. Let us assume the $\hat{\beta}_i$ to be independent of each other for $i = 2, 3, \ldots, 22$. Then, the probability that one of these coefficients at least reaches its critical value, is pretty high, and certainly much higher than 5%. (Indeed, if we make 20 trials of an event which each time has probability 1/20 to happen, the probability that it happens once is then approximately $e^{-1} = 0.36 \neq 0.05$.) So, all of this to justify, why for testing that several coefficients are 0 at the same time, we do not simply use the tests which we would use for each of them individually!

So, again we come back to the situation where we have only the factors weight, size and age. Now, let us first mention that the factors for which we want to test if they have no influence on out target data, must always be placed at the end of the other factors. So, here if we would want to test that the factor weight does not affect milk production, we would have to put first the factors size and age and then after that the factor weight.....
Another reason why we need to test several factors at the same time: imagine again our beloved cows from Switzerland. But say we have 20 factors, many of them quite useless to predict milk production. But for each of them alone when we leave the 19 others it might be difficult to test for its β to be 0 because of colinaerity. So, putting many of them in a package might help a lot, to test that all together they do not contribute!
So, let us explain how we do the testing:

we use the Graham-Schmidt orthogonality applied to the factors

$$\vec{1}, \vec{x}^{weigth}, \vec{x}^{age}, \vec{x}^{size}$$

So, let $\vec{e}_0, \vec{e}_1, \vec{e}_2, \vec{e}_3, \ldots, \vec{e}_{n-1}$ be the sequence of orthogonal vectors obtained from Graham-Schmidt when starting with the sequence: $\vec{1}, \vec{x}^{weight}, \vec{x}^{age}, \vec{x}^{size}$. (We assume that there are n cows in our data set.) This means that $\vec{1}$ and $\vec{e}_0$ are co-linear. Then $\vec{e}_1$ is a linear combination of $\vec{x}^{weight}$ and $\vec{1}$. Similarly $\vec{e}_2$ is a linear combination of $\vec{1}$, $\vec{x}^{weight}$, $\vec{x}^{age}$. Finally, $\vec{e}_3$ is a linear combination of $\vec{1}$, $\vec{x}^{weight}$, $\vec{x}^{age}$ and $\vec{x}^{size}$. After that vectors $\vec{e}_i$ with $i \geq 4$ are simply orthogonal vectors filling the space. We are going to use the two dimensional random vector

$$(\vec{y} \cdot \vec{e}_2, \vec{y} \cdot \vec{e}_3)$$

to base our test for $H_0 : \beta_2 = \beta_3 = 0$. By definition, $\vec{e}_2$ and $\vec{e}_3$ are both orthogonal to $\vec{1}$ and $\vec{x}^{weight}$. Hence, we find

$$\begin{aligned}&\vec{e}_2 \cdot \vec{y} = \vec{e}_2 \cdot (\alpha\vec{1} + \beta_1\vec{x}^{weight} + \beta_2\vec{x}^{age} + \beta_3\vec{x}^{size} + \vec{\epsilon}) = \\ &\vec{e}_2 \cdot (\alpha\vec{1} + \beta_1\vec{x}^{weight}) + \vec{e}_2 \cdot (\beta_2\vec{x}^{age} + \beta_3\vec{x}^{size}) + \vec{e}_2 \cdot \vec{\epsilon} = \\ &= \vec{e}_2 \cdot (\beta_2\vec{x}^{age} + \beta_3\vec{x}^{size}) + \vec{e}_2 \cdot \vec{\epsilon}.\end{aligned}$$

If the hypothesis H_0 is true, then because $\beta_2 = \beta_3 = 0$, the expression above yields:

$$\vec{e}_2 \cdot \vec{y} = \vec{e}_2 \cdot \vec{\epsilon}$$

Similarly for $\vec{e}_3$ we find:

$$\vec{e}_3 \cdot \vec{y} = \vec{e}_3 \cdot (\beta_2\vec{x}^{age} + \beta_3\vec{x}^{size}) + \vec{e}_3 \cdot \vec{\epsilon}$$

which if H_0 is true becomes

$$\vec{e}_3 \cdot \vec{y} = \vec{e}_3\vec{\epsilon}$$

So, if H_0 is true, the expression

$$(\vec{y} \cdot \vec{e}_2, \vec{y} \cdot \vec{e}_3)$$

is equal to

$$(\vec{\epsilon} \cdot \vec{e}_2, \vec{\epsilon} \cdot \vec{e}_3) \tag{5.33}$$

From our Lemma 5.21 on scalar products of non-random vector with $\vec{\epsilon}$, we find that both entries of the vector in 5.33 are independent of each other since $\vec{e}_1$ and $\vec{e}_3$ are orthogonal. They are furthermore both normal with expectation 0 and standard deviation σ. This assuming H_0 to hold. If, H_0 does not hold, then

$$(\vec{y} \cdot \vec{e}_2, \vec{y} \cdot \vec{e}_3) = (\vec{\epsilon} \cdot \vec{e}_2, \vec{\epsilon} \cdot \vec{e}_3) + (\vec{e}_2 \cdot (\beta_2\vec{x}^{age} + \beta_3\vec{x}^{size}), \vec{e}_3 \cdot (\beta_2\vec{x}^{age} + \beta_3\vec{x}^{size})).$$

The difference here to the case of H_0 is the vector

$$(\vec{e}_2 \cdot (\beta_2\vec{x}^{age} + \beta_3\vec{x}^{size}), \vec{e}_3 \cdot (\beta_2\vec{x}^{age} + \beta_3\vec{x}^{size}))$$

which is non-random. Adding a non-random vector changes the expectation but does not change the standard deviation. So basically we are testing for zero expectation in a normal random vector with independent normal entries. We also assume that both entries have standard deviation σ.
Let $N_2 =: \vec{y} \cdot \vec{e}_2$ and let $N_3 := \vec{y} \cdot \vec{e}_3$. Under H_0 thus both N_2 and N_3 have 0 expectation. there standard deviation as mentioned is σ. Now if the expectation is 0 then $VAR[N_i] = E[N_i^2] = \sigma^2$ and hence we would "expect" $\frac{N_2^2+N_3^2}{2\sigma^2}$ to be about 1. We will use $\frac{N_2^2+N_3^2}{2\sigma^2}$ to test the hypothesis $E[N_2] = E[N_2] = 0$ which is equivalent to H_0. This is how we proceed: we find the critical value c_α for the given significance level $\alpha > 0$, so that

$$P(\frac{N_2^2 + N_3^2}{2\sigma^2} \geq c_\alpha) = \alpha. \tag{5.34}$$

The variable N_2 and N_3 being normal with standard deviation σ and expectation 0, we find that equation 5.34 is equivalent to

$$P(\frac{\mathcal{N}_2^2+\mathcal{N}_3^2}{2}\geq c_\alpha)=\alpha$$

where $\mathcal{N}_2$ and $\mathcal{N}_3$ are independent standard normal. This is for the case that we know σ. If we don't know σ, then we simply use $\hat{\sigma}$ instead of σ:
in that case, we use the ratio

$$\frac{N_2^2+N_3^2}{2\hat{\sigma}^2}$$

for our test. If that is bigger than a critical value, then we will reject the hypothesis, otherwise we accept it. Now, remember that our estimate $\hat{\sigma}$ was defined to be equal to

$$\hat{\sigma}^2=\frac{N_4^2+N_5^2+\ldots+N_{n-1}^2}{n-4}$$

where we define N_i to be

$$N_i := \vec{y}\cdot\vec{e}_i$$

for any $i=4,5,6,\ldots,n-1$. We have seen that for i bigger then the biggest index of a factor, we find that $N_i=\vec{\epsilon}\cdot\vec{e}_i$. So, to test if $\beta_2=\beta_3=0$ and assuming we do not know σ, we use

$$\frac{(N_2^2+N_3^2)/2}{N_4^2+N_5^2+\ldots+N_{n-1}^2/(n-4)}=\frac{(\mathcal{N}_2^2+\mathcal{N}_3^2)/2}{(\mathcal{N}_4^2+\mathcal{N}_5^2+\ldots+\mathcal{N}_{n-1}^2)/(n-4)}, \tag{5.35}$$

where $\mathcal{N}_2$,$\mathcal{N}_3$, $\mathcal{N}_4$, ...is a sequence of i.i.d. standard normal. Expression given on the right side of equation **??** is called *a F-statistic* with 2 and $n-4$ degrees of freedom. We can thus look in a table for an F-statistics with 2 and $n-4$ degrees of freedom which is the value c_α which it does not exceed with probability bigger than $\alpha>0$. This is to say that the critical value c_α for our test is given by

$$P(\frac{(\mathcal{N}_2^2+\mathcal{N}_3^2)/2}{(\mathcal{N}_4^2+\mathcal{N}_5^2+\ldots+\mathcal{N}_{n-1}^2)/(n-4)}\geq c_\alpha)=\alpha$$

Once we have determined that critical value for the significance-level α, our test simply consist in checking if the quantity

$$\frac{(N_2^2+N_3^2)/2}{N_4^2+N_5^2+\ldots+N_{n-1}^2/(n-4)}=\frac{(\vec{y}\cdot\vec{e}_2)^2+(\vec{y}\cdot\vec{e}_3)^2}{\hat{\sigma}}$$

exceeds c_α or not. If it does, we reject the hypothesis $H_0=\beta_2=\beta_3=0$ on the level α. Otherwise, we can not reject the hypothesis $\beta_2=\beta_3=0$ on the level α, which could mean several things: the hypothesis is indeed true or there is too much noise in the data for figuring out that the hypothesis is not true. In the second case, adding more cows should after a while help expose the fact that the hypothesis is not true... In general imagine that we have $j+k$ factors and that the model is

$$\vec{y} = \sum_{l=1}^{j+k} \beta_l \cdot \vec{x}^l + \vec{\epsilon}.$$

Again we assume n cows. Say we want to test the hypothesis

$$H_0 : \beta_{j+1} = \beta_{j+2} = \ldots = \beta_{j+k} = 0$$

Then let first $\vec{e}_1, \vec{e}_2, \vec{e}_3, \ldots, \vec{e}_n$ be the orthogonal base obtained when applying Graham-Schmidt to the vector sequence

$$\vec{x}^1, \vec{x}^2, \vec{x}^3, \ldots, \vec{x}^{j+k}.$$

So, we request that the linear subspace generated by $\vec{x}^1, \vec{x}^2, \vec{x}^3, \ldots, \vec{x}^m$ be identical to the subspace generated by

$$\vec{e}_1, \vec{e}_2, \vec{e}_3, \ldots, \vec{e}_m$$

for all $m = 1, 2, \ldots, n+k$. Let $N_i := \vec{y} \cdot \vec{e}_i$ for all $i = 1, 2, \ldots, n$ Then we use as test statistic:

$$\frac{(N_{j+1}^2 + N_{j+2}^3 + \ldots + N_{j+k}^2)/k}{(N_{j+k+1}^2 + N_{j+k+2}^2 + \ldots + N_n^2)/(n-k-j)}. \tag{5.36}$$

which under H_0 has F-distribution with k and $n-j-k$ degrees of freedom. If that statistic is above c_α, then we reject H_0. The critical value for c_α has to be found now in a table for a F-statistics with k and $n-j-k$ degrees of freedom. let us define next SSE and SSR: we define first the sum of squares of errors as:

$$SSE := \sum_{i=1}^{j+k} (y_i - \hat{\beta}_1 x_i^1 - \hat{\beta}_2 x_i^2 - \ldots - \hat{\beta}_i^{j+k})^2$$

Note that our estimate for σ using the full model is simply

$$\hat{\sigma}^2 := \frac{SSE}{n-k-j}$$

So, if SSE is small then this means that when we will have to predict a new cow, typically the estimation error should not be too big. Now, note that SSE is equal to

$$N_{j+k+1}^2 + N_{j+k+2}^2 + \ldots + N_n^2.$$

What is the reason for this? We will see why in the section **??** below. Furthermore we defined the sum of square due to the regression of the full model:

$$SSR(F) := \sum_{i=1}^{n} (\hat{\beta}_1 x_i^1 + \hat{\beta}_2 x_i^2 + \ldots + \hat{\beta}_{j+k} x_i^{j+k})^2$$

and the sum of square of the regression for the restricted model:

$$SSR(R) := \sum_{i=1}^{n}(\hat{\beta}_1^R x_i^1 + \hat{\beta}_2^R x_i^2 + \ldots + \hat{\beta}_j^R x_i^j)^2$$

where

$$\hat{\beta}_1^R, \ldots, \hat{\beta}_j^R$$

denotes the linear regression coefficients in the restricted model where we use only the factors 1 through j instead of 1 through $j+k$. then, the term

$$(N_{j+1}^2 + N_{j+2}^3 + \ldots + N_{j+k}^2)$$

is simply equal to

$$SSR(F) - SSR(R)$$

In other words, if we want to test the hypothesis that $H_0 : \beta_{j+1} = \beta_{j+2} = \ldots = \beta_{j+k} = 0$, then the test statistic given in 5.36 is equal to

$$F = \frac{(SSR(F) - SSR(R))/k}{SSE/(n-j-k)}$$

If the hypothesis H_0 holds, then F has an $\mathcal{F}$ distribution with j and $n-j-k$ degree of freedom as already mentioned.

How to use the statistical data analysis program $\mathcal{R}$ for ANOVA In $\mathcal{R}$, when we use the command

$$fullmodel = lm(\vec{y} \sim \vec{x}^{weigth} + \vec{x}^{age} + \vec{x}^{size})$$

then all the info about the linear regression is stored in the object "fullmodel". Then we do the same for the partial model:

$$partialmodel = lm(\vec{y} \sim \vec{x}^{weigth})$$

The next step is to type into the $\mathcal{R}$-prompt:

$$anova(partialmodel, fullmodel)$$

Then $\mathcal{R}$ gives us a table with the F-statistic and the p-value (among others). What is the p value? Well, we simply compute the numeric value of F and calculate the probability that a random variable which has $\mathcal{F}$ distribution with 2 and $n-4$ degrees be at least that big. So, in other words, the p-value is

$$P(\mathcal{F}_{2,n-4} \geq \frac{(SSR(F) - SSR(R))/k}{SSE/(n-j-k)})$$

where $\mathcal{F}_{2,n-4}$ designates a random variable with an $\mathcal{F}$-distribution with 2 degrees and $n-4$ degrees of freedom. Now note for example, that if the F-statistic takes on the value 19 in our cow data-set, then we know that it is almost impossible that the H_0 be true: if H_0 is true, then the statistic

$$F = \frac{(SSR(F) - SSR(R))/k}{SSE/(n-j-k)})$$

is equal to an average of standard normal squared divided by an average of other standard normal squared. A standard normal squared has an expectation of 1. So, an average of standard normal squared we should expect to be about 1 by the law of large numbers. hence, 1 over 1 should be about 1 and not 19. So, with the F-statistic being 19, it is almost impossible that H_0 be true. But how "impossible" is it? for this we look in the table of $\mathcal{F}$-statistics. In this case with 100 cows, we get that the second degree of freedom would be 96. So, with 2 and 96 degrees of freedom the probability for an F-statistic to exceed 19 is 0.00000012. this would then be the p-value in our case. This is simply so small, that it will not happen in practice! Hence, we are sure that the hypothesis would not be true, that is the factors size and age add predictive power to the factor weight.

5.7 SSE and SSR

let us consider a situation where we want to predict milk with m factors:

$$Y_i = \beta_1 x_i^1 + \beta_2 x_i^2 + \ldots + \beta_k x_i^m + \epsilon_i$$

where Y_i denotes the amount of milk produced by the i-th cow. x_i^l is the value for the l factor in cow number i for all $i = 1, 2, \ldots, m$. (this means that if for example the l-th factor is weight, then x_i^l is the weight of the i-th cow). Again, ϵ_i denotes the i-th cows individual variation term. Let $\vec{x}^l$ denote the l factor vector:

$$\vec{x}^l := (x_1^l, x_2^l, \ldots, x_n^l)^t$$

where we assumed that there are n cows total. If $\vec{Y}$ is the column vector with the milk, then our model reads:

$$\vec{Y} = \beta_1 \vec{x}^1 + \beta_2 \vec{x}^2 + \ldots + \beta_m \vec{x}^m + \vec{\epsilon}.$$

The equations to determine the estimate for the correlation coefficients β_l are given by

$$0 = \sum_{i=1}^{m} (Y_i - \hat{\beta}_1 x_i^1 - \hat{\beta}_2 x_i^2 - \ldots - \hat{\beta}_i^m \vec{x} + i^m) \cdot x_i^l = (\vec{Y} - \hat{\beta}_1 \vec{x}^1 + \hat{\beta}_2 \vec{x}^2 + \ldots + \hat{\beta}_m \vec{x}^m +) \cdot \vec{x}^l$$

where we have one such equation for every $l = 1, 2, \ldots, m$. Hence $\vec{x}^l$ and

$$(\vec{Y} - \hat{\beta}_1 \vec{x}^1 - \hat{\beta}_2 \vec{x}^2 - \ldots - \hat{\beta}_m \vec{x}^m)$$

are orthogonal for all $l = 1, 2, \ldots, m$. Thus

$$\hat{\beta}_1\vec{x}^1 - \hat{\beta}_2\vec{x}^2 - \ldots - \hat{\beta}_m\vec{x}^m$$

and

$$(\vec{Y} - \hat{\beta}_1\vec{x}^1 - \hat{\beta}_2\vec{x}^2 - \ldots - \hat{\beta}_m\vec{x}^m)$$

are orthogonal to each other. But, their sum is $\vec{Y}$. When we have a sum of orthogonal vectors we can apply Pythagoras. Hence the length square of $\vec{Y}$ must be equal to the sum of the length squares of the two vectors:

$$|\vec{Y}|^2 = |(\vec{Y} - \hat{\beta}_1\vec{x}^1 - \hat{\beta}_2\vec{x}^2 - \ldots - \hat{\beta}_m\vec{x}^m)|^2 + |\hat{\beta}_1\vec{x}^1 - \hat{\beta}_2\vec{x}^2 - \ldots - \hat{\beta}_m\vec{x}^m)|^2. \tag{5.37}$$

Note that the two parts in the sum on the right side of the last equation above are just the sum of square of the errors (SSE) and the sum of square of the regression (SSE) of our model:

$$(\vec{Y} - \hat{\beta}_1\vec{x}^1 - \hat{\beta}_2\vec{x}^2 - \ldots - \hat{\beta}_m\vec{x}^m)|^2 = \sum_{i=1}^{m}(Y_i - \hat{\beta}_1 x_i^1 + \hat{\beta}_2 x_i^2 + \ldots + \hat{\beta}_m x_i^m)^2 = SSE$$

It should also be clear why this is called SS-error: if we were to predict the milk of a cow Nb i we would use the formula

$$\hat{\beta}_1 x_i^1 + \hat{\beta}_2 x_i^2 + \ldots + \hat{\beta}_m x_i^m.$$

the milk produced by that cow is Y_i so the estimation error is

$$\hat{\epsilon}_i := Y_i - (\hat{\beta}_1 x_i^1 + \hat{\beta}_2 x_i^2 + \ldots + \hat{\beta}_m x_i^m.)$$

And, the SSE is just the sum of these estimation errors squared:

$$SSE = \sum_{i=1}^{n} \hat{\epsilon}_i^2.$$

Now, we don't need to predict how much milk cow number i produces, because we already have that information in our data set. But still, to see how good our milk prediction technique we look how good it would be for the given cows in our data-set:
for each cow in the data-set we calculate the estimation error and then build the average of the estimation errors squares. This should give a good idea of what the average estimation error square is also for new cows.
Now, let $\vec{e}_1, \vec{e}_2, \ldots, \vec{e}_n$ be the orthogonal basis which we obtain, when we apply Graham-Schmidt normalization to the factor vector sequence:

$$\vec{x}^1, \vec{x}^2, \vec{x}^3, \ldots, \vec{x}^m$$

Hence, $\vec{x}^1$ and $\vec{e}_1$ are co-linear. Then, $\vec{x}^1$ and $\vec{x}^2$ generate the same vector space as $\vec{e}_1$ and $\vec{e}_2$. Similarly for any $l \leq m$, we have that the set

$$\vec{x}^1, \vec{x}^2, \ldots, \vec{x}^l$$

generates the same vector subspace as

$$\vec{e^1}, \vec{e^2}, \ldots, \vec{e^l}.$$

Anyhow, we have that $\vec{e}_1, \vec{e}_2, \ldots, \vec{e}_3$ is a orthonormal basis of our vector space. Thus for any vector

$$\vec{a} = (a_1, a_2, \ldots, a_n)^t$$

in our vector space, we have that $\vec{a}$ can be represented in our basis by:

$$\vec{a} = (\vec{a} \cdot \vec{e}_1)\vec{e}_1 + (\vec{a} \cdot \vec{e}_2) \cdot \vec{e}_2 + \ldots + (\vec{a} \cdot \vec{e}_n) \cdot \vec{e}_n$$

and the length square of the vector $\vec{a}$ is the sum of the squares

$$|\vec{a}|^2 = \sum_{i=1}^{n} a_i^n = (\vec{a} \cdot \vec{e}_1)^2 + (\vec{a} \cdot \vec{e}_2)^2 + \ldots + (\vec{a} \cdot \vec{e}_n)^2.$$

To, if we apply this talking the vector $\vec{a}$ equal to:

$$\vec{a} = \vec{Y} - \hat{\beta}_1\vec{x^1} - \hat{\beta}_2\vec{x^2} - \ldots - \hat{\beta}_m\vec{x^m})$$

, then we have

$$SSE = |\vec{a}|^2 = \sum_{j=1}^{n} (\vec{e}_j \cdot \vec{a})^2. \tag{5.38}$$

But now if $j > m$, then $\vec{e}_j$ is by definition perpendicular to any factor vector $\vec{x^l}$, with $l \leq m$. so, then, in the case $j > m$ we have:

$$\vec{e}_j \cdot \vec{a} = \vec{e}_j \cdot (\vec{Y} - \hat{\beta}_1\vec{x^1} - \hat{\beta}_2\vec{x^2} - \ldots - \hat{\beta}_m\vec{x^m}) = \vec{e}_j \cdot \vec{Y}$$

On the other hand if $j \leq m$ we have that $\vec{e}_j$ is in the same space as the space generated by the first j factors. but the factors are as we saw all perpendicular to $\vec{a}$, and hence $\vec{e}_j$ must also be perpendicular to $\vec{a}$. This is to say, that in that case that $j \leq m$, we have

$$\vec{e}_j \cdot \vec{a} = 0$$

Summarizing what we found for $\vec{e}_j \cdot \vec{a}$, and using it with equation 5.38, we find

$$SSE = \sum_{j=m}^{n} (\vec{Y} \cdot \vec{e}_j)^2$$

But recall that our unbiased estimate for σ^2 is given by

$$\hat{\sigma}^2 = \frac{\sum_{j=m}^{n} (\vec{Y} \cdot \vec{e}_j)^2}{n - m}$$

and hence we find that our unbiased estimate for the variance is equal to

$$\hat{\sigma} = \frac{SSE}{n-k}.$$

Often time the sum $\sum_{i=1}^{n} Y_i^2$ is denoted by SST where T stands for "total". So

$$SST := \sum_{i=1}^{n} Y_i^2.$$

Now, with that definition, equation 5.37 reads:

$$SST = SSR + SSE \tag{5.39}$$

5.8 Finding the best model for prediction

When we consider the same data set and a bunch of factors given in the data set, we might still not want to use all factors for our prediction. the reason is as follows: assume that we have as factors for the milk as usual weight, size and age and then 20 others. Assume that the true β coefficient of all the 20 other factors is 0. So, the true model would be

$$\vec{y}_i = \alpha + \beta_1 x_i^{weight} + \beta_2 x_i^{size} + \beta_3 x_i^{age} + \epsilon_i$$

So, we might not know that and use for prediction all 20 factors, and hence predict the amount of milk a new cow produces to be

$$\hat{\alpha} + \hat{\beta}_1 x_{new}^{weight} + \hat{\beta}_2 x_{new}^{size} + \hat{\beta}_3 x_{new}^{age} + \hat{\beta}_3 x_{new}^{4} + \hat{\beta}_4 x_{new}^{4} + \ldots + \hat{\beta}_{22} x_{new}^{22}.$$

Now if really we would have

$$\beta_4 = \beta_5 = \ldots + \beta_{22} = 0 \tag{5.40}$$

or close to that situation, then the term

$$\hat{\beta}_3 x_{new}^{4} + \hat{\beta}_4 x_{new}^{4} + \ldots + \hat{\beta}_{22} x_{new}^{22} \tag{5.41}$$

would be superfluous and may even be harmful. Actually then that additional term, if there is a lot of co-linearity could have big estimation errors for the $\hat{\beta}_i$'s for which $i \geq 4$. but assuming the true β_i's to be zero, this could mean that the $\hat{\beta}_i$'s for $i \geq 4$ could be far away from zero, making thus the expression 5.41 really big. So, we would have one additional "noise" part which could very much reduce the precision of our estimate! To remedy this, we could make a test to figure out if **??**, and if we find that it might, we would then work with the reduced model. All this to say that if we have a lot of factors, it is usually not good to all use them for prediction. So, one has to select which ones one wants. This is called *model selection.* Which model do we select? Well we want the prediction to be as precise as possible. So, we could chose the one model which has the smallest unbiased estimate for σ. Let us explain: let us consider three models:

$$\begin{aligned}
\texttt{Model I :}\vec{Y} &= \alpha^{I}\vec{1} + \beta_1^{I}\vec{x}^{weight} + \beta_2^{I}\vec{x}^{age} + \vec{\epsilon} \\
\texttt{Model II :}\vec{Y} &= \alpha^{II}\vec{1} + \beta_1^{II}\vec{x}^{weight} + \vec{\epsilon} \\
\texttt{Model III :}\vec{Y} &= \alpha^{III}\vec{1} + \beta_1^{III}\vec{x}^{weight} + \beta_2^{III}\vec{x}^{age} + \vec{\epsilon}
\end{aligned}$$

Note that when the factors are not orthogonal, then the coefficients and their estimates will depend on which model we consider. This is why, we wrote the coefficients with a superscript indicating the model, they are calculated for.

How should we now select the model which is best suited for prediction among the three models above? (Assuming we want to use one of those three models). Again, if we put in factors which in reality have a β-coefficient equal to 0, then this can only make the estimation process worse. Let us consider the situation where Model I would be the true model. Say we work a lot with the model and know the coefficients exactly. Then the 95%-confidence interval for the amount of milk of a new cow would be

$$\alpha^{I} + \beta_1^{I}x_{new}^{weight} + \beta_2^{I}x_{new}^{age} \pm 1.96\sigma_I$$

or we can write it as

$$[\alpha^{I} + \beta_1^{I}x_{new}^{weight} + \beta_2^{I}x_{new}^{age} - 1.96\sigma_I, \alpha^{I} + \beta_1^{I}x_{new}^{weight} + \beta_2^{I}x_{new}^{age} + 1.96\sigma_I].$$

So, the level of precision here is given by σ_I. So, we could chose which model is best for prediction, but taking the one for which σ is smallest. Hence, we would compare σ_I, σ_{II} and σ_{III} and chose the model for which that value is smallest. Now, note in most cases we might not know exactly what α^I, β^I and β^I are. So, we will use the estimates instead. This will make the imprecision for our confidence interval slightly bigger. But, in general, except with very big co linearity, the order of the size of the confidence interval will remain the same. So, we can still take σ_I has a measure of approximately how precise Model I is for prediction. This can also be seen, in our formula **??** for the precision of estimates for the coefficients: mostly we get stuff of the type

$$\frac{\sigma}{\sqrt{n}}$$

times constant for the precision of $\hat{\beta} - \beta$. So, because of the $\sqrt{n}$ in the denominator, this then becomes much smaller than σ and hence σ still gives the approximate order of the prediction precision even if we have to estimate the correlation coefficients. Often of course the value of σ is not known. So, instead we will use its estimate. Then we just chose the model for which the estimate of σ is smallest. We use the unbiased estimate. So,

$$\hat{\sigma}_I^2 := \sum_{i=1}^{n} \frac{(Y_i - \alpha^I + \beta^I x_i^{weight} + \beta^I x_i^{age})^2}{df_I}$$

where *df* designates the degree of freedom here $df_I = n-3$. then we calculate the estimate for σ in the second model:

$$\hat{\sigma}_I^2 := \sum_{i=1}^{n} \frac{(Y_i - \alpha^{II} - \beta_1^{II} x_i^{weight})^2}{df_{II}},$$

where the degree of freedom is $df_{II} = n-2$ finally we calculate the estimate for σ according to the third model:

$$\hat{\sigma}_I^2 := \sum_{i=1}^{n} \frac{(Y_i - \alpha^{III} - \beta_1^{III} x_i^{weight} - \beta_2^{III} x_i^{age})^2}{df_{III}},$$

with $df_{III} = n-3$. then find the smallest among: σ_I^2, σ_{II}^2 and σ_{III}^2 and chose the corresponding model.
What about if we have a lot of factors? Say we would have for example 30 factors. Then the number of possible models obtained from including several of these factors is equal to $2^{30} \geq 1000000000$!! so, that makes it difficult to calculated. Also, it will create a big problem of the type over fitting or you could call it "data-snooping". What does that mean? If you have that many models to compare and for each model M you have an estimate $\hat{\sigma}_M$, then by the large number you will have some where the estimate $\hat{\sigma}_M$ is very wrong and much smaller than the actual σ_M. If we would know the true σ_M for each model M, then we would simply select the model for which σ_M is smallest. But, as mentioned, we don't know the true value of σ_M but only have an estimate (approximation) of it. This estimate is quite fine usually, but with a billion of Models its just bound to have some of these estimates which will be very wrong! So, when selecting a model if we would chose among all one billion models and pick the one with smallest $\hat{\sigma}_M$ here is what will happen: the chances are very high that we pick a model M_0 which does not have its true σ_{M_0} small, but instead has just a small $\hat{\sigma}_{M_0}$ dues to a big estimation error! With a billion of estimates, their is bound to be a few with a quite substantial error in the estimate. (Here we look for an error which would underestimate the true value).
This is the same phenomena, as if you would data-mine stocks of firms until you find something which is correlated substantially to an above average return. if you work for days, and look at millions of parameters, you may find in the end for example, that the firms with a CEO who has dark hair, and CFO with a name which starts in M, outperforms other stocks in a significant way. Then we try to apply this what we found by investing only into those firms where CEO has dark hair and CFO has a name starting with M. Of course, we will probably not get an above average return with such a policy: the reason is that we looked for millions of things, and in the end found something which by chance was correlated to above expected return. But then in the future this will not be the case again!
To solve this problem in the case there are many factors, there are different approaches possible. The one which makes most sense to us goes as follows:
First chose among all models with only one factor the one which has smallest $\hat{\sigma}^2$. Say you find that the one factor which is best suited is $\vec{x}^{weight}$. Then, in the second step, you

consider all two factor models where one of the factors is $\vec{x}^{weight}$. From those two factor models you chose again the one which minimizes $\hat{\sigma}^2_M$. Say you find that at that second level the best is to use $\vec{x}^{weight}$ with $\vec{x}^{age}$. So, for the third step you consider only three factor models which contain $\vec{x}^{weight}$ and $\vec{x}^{age}$ and any third factor available. You chose then among those three factor-models the one which minimizes $\hat{\sigma}^2_M$. You go on adding one factor at a time to the preciously chosen ones. In this manner with n factors you get only $n^2/2$ choices (polynomial number in n) instead of the exponentially big number 2^n. Another improvement you can make is as follows: you keep on adding factors, but you stop as soon as you realize that adding more factors does't improve things much. One way to do this, is at each step to test for the remaining factors to have all coefficients equal to 0 on a given significance level. Another approach is to check by how much $\hat{\sigma}_m$ gets improved. If it is not improving a lot from one step to the next then stop the procedure and stay with the factors you found for your model selection. So for example you could say that from one step to the other the estimate $\hat{\sigma}$ has to improve by at least 20%. Say M_i is the model selected at the i-th step and M_{i+1} is the model selected at the $i+1$ step. Then you could say that you go on until for example the following is no longer true:

$$\frac{\hat{\sigma}_{M_{i+1}}}{\hat{\sigma}_{M_i}} \leq 0.8$$

Another thing to realize is that we chose among different models which model M among a set of different one has smallest $\hat{\sigma}^2_M$, then this is the same as finding the one which has biggest $SSR(M)$. Indeed we have seen in 5.39, that

$$SST = SSR(M) + SSE(M).$$

But, SST does not vary as we change the model, since it only depends on the milk produced by each cow. So, minimizing $SSE(M)$ is the same as maximizing $SSR(M)$. Now, our estimate is not exactly $SSE(M)$ but is the faction:

$$\hat{\sigma}^2_M = \frac{SSE(M)}{df_M} \tag{5.42}$$

where df_M is the degree of freedom of the model, which is equal to the number of cows n minus the number of factors in the model. Since at each step we compare only models with the same number of factors, when we minimize 5.42, this is the same as minimizing $SSE(M)$ since then df_M is the same for all models we compare at the same time. Hence, at each step what we do is equivalent to minimizing $SSE(M)$, which by equation 5.39 amounts to the same as maximizing $SSR(M)$. I only mention this so you know why in tables printed out by $\mathcal{R}$, they always give $SSR(M)$.

One more approach would be to keep the data of a few randomly selected cows just to decide at which moment to stop our procedure. That is we could estimate $\hat{\sigma}_M$ as described until now. Then at each step we would also re-estimate it by using the additional cows, but using the parameters $\hat{\beta}_i$ found from the big data-set of cows. In this way, we would not have the over-fitting problem for this part of the best model search....

5.9 General overview

In a somewhat oversimplified way, we could say there are two somewhat different problems at hand:

1. Prediction. For example this could be: We are going to build a new shop. We want to estimate the future profit and give a confidence interval. We are a big company, so we have a lot of data about our shops in the US and their profits. We can put y_i to be the profit the i-th shop makes. As factors we could use for each shop's neighborhood the following: the population density, the average income, percentages of age groups, number of other shops in area, other indicators of economical situation in area, presence of other competing shops, part of a Mall or not, business of street next to shop,.....So, we know all the factors for the new shop and can use a multidimensional regression model to predict the profit. We will also give a confidence interval for the profit. We don't care so much if some of the β_i's are wrong, as long as the overall estimate of the profit is pretty correct. This means that often for this type of prediction problem, the co-linearity is not such a severe problem: when we select a model, we ill simply not chose factors to put into the model which would have a lot of co-linearity between them. Say, back to the milk and cow problem. Probably size and weight are similar in the sense that bigger cows tend to be heavier. so, these two factors are very much correlated to each other. but typically the model selection which we proposed in the previous paragraph, will then only select one of these two to be included into the model: say your best single factor is weight. Then the second which you chose will probability be age and not size: size is similar in its information content to weight, so it will probably work better to include age at the second step rather than size. So, this shows that with the model selection procedure explained in the last subsection, there is a tendency to not include highly similar factors, that is factors with a lot of co-linearity. So, typically if we would include all the factors we may have a lot of co-linearity, but the model we chose will often not have that problem.

2. Testing for a given i if $H_0 : \beta_i = 0$. An example for this is a long term study about cardiovascular health. Here Y_i could be a score of individual number i in the study, which expresses the cardiovascular health of that individual. Then in principle, the co-linearity problem can not just be ignored and close factors thrown out of the model! let us explain why: in prediction for cows for example, size and weight might be very similar for the purpose of predicting milk production. So, in the model selection part we might chose either one of them (will be a little bit of chance which one gets selected). Then with each of them we would probably achieve similar precision for estimating milk production. So, we don't care if in the model selection one gets replaced by the other. This might happen because of co-linearity. Now, for a problem like cardiovascular health, this is very different: say virgin olive oil with the omega 3 is very good on the long run for your hart. But spaghetti are just neutral. Now, in Italy, in many regions, (at least when I

was a kid) they ate paste almost every day. Back then olive oil was used mainly in southern countries in Europe and not in Northern Europe. So, I would expect, that 30 years ago in Europe, eating a lot of pasta and consumption of olive oil, were very much correlated. So, you would have co-linearity between these two factors. but now assume that you have found that to predict cardiovascular health spaghetti eating is the best predictor. In reality it is the olive oil which causes the better health. So, for prediction this is fine, but not for understanding what is going on! Say you are an insurance company who has to decide how much a person has to premium to pay. if you can predict cardiovascular health well, you are in business even if you use the wrong factor "spaghetti" instead of the true cause of health which is "Omega 3 in olive oil" But for a doctor, which is investigating how to improve the health of people, thinking that spaghetti is good for your hart is totally wrong: say a government then promotes spaghetti eating as a public health approach to improving cardio-vascular health. That would then not be helpful. So, when we want to find if $\beta_{olive\ oil}$ is 0 or not, we can not simply discard other factors which are closely co-linear to $\vec{x}^{olive\ oil}$. This is why these type of long term studies are so difficult. In principle, the only way is then to increase the number n. That is we have to include more people in our study. As a matter of fact, the precision of our estimate for β_{olive_oil} is given by something of the type

$$\frac{\sigma}{\sqrt{n} \cdot sd(\vec{x}^{olive\ oil}\delta}$$

So, the precision gets as good as we want in principle when n becomes big enough. Hence, we will be able to figure out if the true $\beta_{olive\ oil}$ is close to 0 or not by just increasing n, that is adding more people to the study. Which is why these type of studies need to be on very large scale and are often very difficult. Of course, factors which are very "far" from $\vec{x}^{olive\ oil}$ and which have a lot of co-linearity between themselves but not with $\vec{x}^{olive\ oil}$ may be kicked out. Then again that might not help that much, because of how we calculated δ: the distance from the space generated by the other factors to the re-scaled $\vec{x}^{olive\ oil}$. Now, those factors which anyhow are far away for $\vec{x}^{olive\ oil}$ may not effect δ all that much. Another important problem is that we are interested in this type of problem in causality: we don't just want to know if things are correlated within our model (that is $\beta_{olive\ oil} > 0$) but if truly *oliveoil* is the true reason for better cardiovascular health. Now if you forget to include one factor, then this might cause you to see that a coefficient is positive, though it is not the cause for better health: say again olive oil is causing better health, but this factor has not been included in your data set. Instead you have include sardines as a possible factor in your list. Then, though the true reason for better cardiovascular health is olive oil, but since this factor is not available in your data set, linear regression analysis jumps to the conclusion that sardines is what creates better cardio-vascular health. The reason is that the two are very correlated, because again in the south of Europe they eat sardines as a main food all the time and that is also where the main olive oil consumption takes place. The only way around this is to try to check for all possible factors. Again, for stuff like cancer, where you have maybe million of chemicals in our food and a few of them could have an effect of increasing the likelihood is very difficult!

5.10 Facing the real data

We have a nice theory, but it is based on assumption which are difficult to check in practice. Among others, independence is often almost impossible to check. Now, what if, all the nice assumptions, like normality, independence and so on would not be there, would our approach still work? The answer is in many cases, linear regression still works. First note that for the estimates $\hat{\alpha}$, $\hat{\beta}_1$ and $\hat{\beta}_2$ to be unbiased we only need the error terms to have zero expectation and nothing else. (Go check the proof). So, But, 0 expectation we can get just by definition: instead of having the model

$$Y_i = \alpha + \beta_1 x_i^{weight} + \beta_2 x_i^{age} + \epsilon_i,$$

define

$$\epsilon_i := Y_i - E[Y|x_i^{weight}, x_i^{age}]. \tag{5.43}$$

So, here ϵ_i is the difference between the amount of milk the i-th cow truly produces and what we would expect given the weight and age. With this definition we get

$$\begin{aligned} &E[\epsilon_i] = E[Y_i - E[Y|x_i^{weight}, x_i^{age}]]\ ] = \\ &E[Y_i] - E\ E[Y|x_i^{weight}, x_i^{age}]\ = E[Y_i] - E[Y_i] = 0 \end{aligned}$$

Now let $g(x^{weight}, x^{age})E[Y|x^{weight}, x^{age}]$ be the function which gives the expected amount of milk given weight and age. Then, by definition

$$Y_i = g(x^{weight}, x^{age}) + \epsilon_i$$

where ϵ_i is defined by **??**. Now, $g(.,.)$ might not be a simple linear function, but it is probably smooth. And hence it can be approximated by a polynomial in the factors as closely as we want. (This is the result of the multidimensional Taylor theorem). Say we find that $g(x^{weight}, x^{age})$ can be approximated sufficiently precisely for our purpose by a second order polynomial:

$$g(x^{weight}, x^{age}) \approx \alpha + \beta_1 x^{weight} + \beta_2 x^{age} + \beta_3 (x^{weigth})^2 + \beta_4 (x^{weight} \cdot x^{age}) + \beta_5 (x^{age})^2$$

then the linear model is:

$$\vec{Y} = \alpha\vec{1} + \beta_1 \vec{x}^{weight} + \beta_2 \vec{x}^{age} + \beta_3 \vec{x}^{weigth\cdot weight} + \beta_4 \vec{x}^{weight\cdot age} + \beta_5 \vec{x}^{age\cdot age} + \vec{\epsilon} \tag{5.44}$$

where

$$\vec{x}^{weigth\cdot weight} = (x_1^{weight} \cdot x_1^{weight}, x_2^{weight} \cdot x_2^{weight}, \dots, x_n^{weight} \cdot x_n^{weight}),$$

and

$$\vec{x}^{weigth\cdot age} = (x_1^{weight} \cdot x_1^{age}, x_2^{weight} \cdot x_2^{age}, \dots, x_n^{weight} \cdot x_n^{age}),$$

and finally

$$\vec{x}^{age \cdot age} = (x_1^{age} \cdot x_1^{age}, x_2^{age} \cdot x_2^{age}, \ldots, x_n^{age} \cdot x_n^{age}),$$

So, if we are willing to sacrifice the idea that the expected among of milk is a linear function of weight and age, we get automatically, that the error terms ϵ_i have expectation 0, and this guaranties at least that the estimates $\hat{\alpha}$, $\hat{\beta}_1$ and $\hat{\beta}_2$ are unbiased! In terms, of the vector model, like in 5.44, we still get a linear vector model even if the function $g(weight, age) = E[milk|weight, age]$ is not linear in weight and age, but is just a finite polynomial in weight and age. so, all our techniques apply, except that we will have more factors!

Now, what about the precision of our prediction for a new cow? So, first note that if we would just have uncorrelated errors each having same standard deviation, then all our calculations about standard deviation of estimates $\hat{\alpha}$,$\hat{\beta}_1$ and $\hat{\beta}_2$ remain valid.
But even that in reality might not be exactly the case: some cows might be a little bit correlated. Also, they might not all have exactly the same σ. But here is an argument, which explains why we are often able to use linear regression in situation where it is not exactly true that we have all the error terms are uncorrelated and have exactly same standard deviation: Say the new cow you want to predict, is “similar” to the others. When you add her to the data set and calculate our estimates all the estimates remain about the same. Then we could imagine the following “Gedankenexperiment”: you add her, but she is similar, so after adding her we could pick at random any of the cows and it “would somehow amount to the same”. that would tell you that the expected error square should be about

$$\frac{\sum_{i=1}^{n+1}(Y_i - \hat{\alpha} - \hat{\beta}_1 x_i^{weight} - \hat{\beta}_2 x^{age})^2}{n+1}$$

which is about our Maximum-Likelihood estimate for σ. (Except that we have an added cow).....So, this argument, tells me that things might work, as long as the cow is “typical”, in the sense that she is not drawn from another population then the ones in the data-set......

6 Invariant testing, principal components and the mutivariate T^2 test

First assume that you have an artillery gun g_0 shooting without changing its position, amunition or direction and/or elevation. Let X, resp. Y be the x-coordinate, resp y-coordinate of the impact point. Assume at first for simplificaton that $E[X] = E[Y] = 0$ and that X and Y are independent of each other with $\sigma_X = \sigma_Y = 1$. Now, we observe suddenly an impact point $x = 2.3$ and $y = 0.3$. This seems a little bit far from our expected impact point to come from our artillery gun. maybe it is another gun, shooting which is also execising in the same area. We would like to make a statistical test on the 5% confidence level. We assume that the other gun has same covariance matrix. In the current case the covariance matrix is the identity. Now, for one dimensionl normal variable, we are with 95% probability withing 1.96 the standard deviation from the expected value. So, if we look at the x coordinate we are further: we are 2.3 standard deviations away from the expectation of gun g_0. Hence, if we would be given only $x = 2.3$ as information and not the other coordinate, then we would reject the hypothesis that it is gun g_0 shooting. But, when we use both coordinates, we use $X^2 + Y^2$ as test statistic. (We look at the distance square of the impact point from the expected value of our gun g_0. If the distance square is too big we reject the hypotesis that it is our gun, as simple as that). Now, if it is our gun shooting then $X^2 + Y^2$ is a sum of two independent standard normals which got squared. (It is the distance square from the origin of a normal vector with indpendent standard normal entries). In statistics such a variable is called Chi-square with 2 degrees of freedom. So, we go into a table and look for the critical value at 5%s significance. We find 5.99. This is to say that the sum of two independent standard normals-squared exceeds 5.99 only in about 5% of cases. In our case, using Euclide, the distance square to the origin of this new impact point is $x^2 + y^2 = 2.3^2 + 0.3^2 = 5.38$ is below being significant. **So, in the two dimensional case, we can not reject the new impact point as being from our original artillery gun on the 5%-level, despite a test on one coordinate only would give a different result**. What is the reason for this difference? Well imagine a high dimensional data, with lots of coordinates which are independent of each other, not just two. Then, because there are lots of coordinates, there is a high probability that at least some of them will be further away from their expectatin than that is typical. (Of course, most will jsut behave in a regular was, and be within two standard devition from their mean). Hence, we can not just test every coordinate individually, because some will tell us to reject the hypothesis. Simple, because in a large enough group there will always be a few outliers. So, to take this into account we take the average of the squares.

Now let us go more formal: consider testing for a normal random vector (X, Y) with independent standard normal entries: the hypothesis $E[X] = E[Y] = 0$ (that is $\mu_x, \mu_y = (0, 0)$ against $(\mu_x, \mu_Y) = (E[X], E[Y]) \neq (0, 0)$. (So, both hypothesis have the same covariance matrix equal to I.) We assume that we know nothing about what (μ_x, μ_y) the average impact point looks like if it is another gun shooting. (That is if we are in the alternative hypothesis.) So, if for example there are two batteries shooting and we know the approximate position of the other battery, and the alternative would be that the other battey has been shotting and we know approximately there average impact point, then we would have a totaly different situation. That is we would test our gun against a gun of the other battery. That would mean test average impact point $(0, 0)$ against another given, known impact point (μ_x, μ_y). Same covariance. We have seen then the optimal deicision rule is a half space. Totally different stuff from using the distance square to the origin....)

What is the justification for using the distance square to the averge impact point of our gun as test statistic? One way of justifying it, is that the circle is invariant under rotation. So, when we rotate the data around $(0, 0)$ we transform the problem into itself. The distributions in the alternative hypothesis are changed, individually, but not as a set. So, since the problem remains invariant under rotation around the average impact point of g_0, we may want the acceptance region to satisfy the same invariance. This allows in many situations to find a way of testing: request the acceptance region to be invariant under a set of transformations which leave the problem invariant. **This is a very abstract way, from the cook-book of abstract math-statistics to justify the circle around the average impact points as acceptance region**

Now, most artillery guns have different latteral flucutation than fluctuation in the direction of shooting. Imagine, now that X and Y are still independent but have different standard deviations. Say you have $\sigma_X = 1$ and $\sigma_Y = 3$. for our artillery gun g_0, where still $E[X] = E[Y] = 0$. We get an impact point $(x, y) = (1.9, 5.9)$. Note that each coordinate separately is within 2 standard deviation from its expectatin. (Provided we shoot with our gun g_0). So, if we do tests for the coordinates separately we would not reject that the shell comes from the gun g_0). But, let us do a two dimensional test. For this we divide the coordinates by their standard deviation. Then, this being done we have two independent standard normals. So, we are back in the previous situation, where we can use the Chi-square statistic for a test. In other words, we take

$$\frac{x^2}{1^2} + \frac{y^2}{3^2} = \frac{1.9^2}{1} + \frac{5.9^2}{9} = 7.4\bar{7} \tag{6.1}$$

as testing statistics. This is far from the critical value at 5% significance. Indeed, we had seen that value being at 5.99. Hence, working in two dimension, we have significant evidence (on the 5% level) that it is not our gun g_0 which has shot the shell with impact point $(1.9, 5.9)$. This would not have been detected if we had just looked at the coordinates individually!!

What is the justification for taking the expression 6.1 for our testing purpose as test statistic? One can show again that the acceptance region given by

$$\frac{x^2}{1^2} + \frac{y^2}{3^2} \leq \texttt{constant} \tag{6.2}$$

are the only possible ones if we want invariance under all linear transforms which leave the probem invariant. (That is all linear maps which leave the covariance matrix invariant. So, we would have to find all 2×2 matrices A for which $A^t COV A = COV$, where COV is the covariance matrix of (X, Y). Then we would need to determine all sets which are invariant for all such A's). That would be a lot of work. So, instead we transform our impact point into having standard normal independent coordinates and then apply the invariance principle. In other words, we apply the invariance principle to $(X/\sigma_X, Y/\sigma_Y)$. There, we have independent standard normal entries, where the invariant acceptance region is the circle. so, the optimal test has acceptance region:

$$\frac{X^2}{\sigma_X^2} + \frac{Y^2}{\sigma_y^2} \leq \texttt{constant},$$

where the constant depends on the significant level one wants.

Next situation is when we have again an impact point, but the coordinates are not independent of each other. Then, we can transform the problem into the same situation that we had previously by using the *principal components.* So, let the impact point be the random vector $\vec{X} = (X,Y)^t$. The principal components $\vec{\nu}_1$ and $\vec{\nu}_2$ are the eigenvectors of the covariance matrix. We take them to be of length 1. The corresponding eigenvalues are designated by λ_1 and λ_2. Let us cite a few facts from Matzingers lecture notes on machinelearning about principal components:

1. The principal components $\vec{\nu}_1$ $\vec{\nu}_2$ are orthogonal to each other.

2. If we express $\vec{X} = (X,Y)$ in the coordonate system defined by $\vec{\nu}_1$ and $\vec{\nu}_2$, then the new coordonates are independent of each other. That is $\vec{\nu} \cdot \vec{X}$ and $\vec{\nu}_2 \cdot \vec{Y}$ are independent of each other.

3. Using the principal components as coordinate system, the variances of the coordinates are the eigenvalues. So $\lambda_1 = VAR[\vec{\nu}_1 \cdot \vec{X}]$ and $\lambda_2 = VAR[\vec{\nu}_2 \cdot \vec{X}]$.

So, since in the coordinate system of the principal components, the coordinates are independent, we can just apply what we have seen so far and find as acceptance region:

$$\frac{X_{PC}^2}{\sigma_{X_{PC}}^2} + \frac{Y_{PC}^2}{\sigma_{Y_{PC}}^2} \leq \texttt{constant} \tag{6.3}$$

where $X_{PC} = \vec{X} \cdot \vec{\nu}_1$ and $Y_{PC} := \vec{X} \cdot \vec{\nu}_2$ are the coordinates in the principal component coordinate system and the constant dependents on the significance level we want. So, let us give an example. In other words, the acceptance region can be written as

$$\frac{(\vec{X} \cdot \vec{\nu}_1)^2}{\lambda_1} + \frac{(\vec{X} \cdot \vec{\nu}_2)^2}{\lambda_2} \leq \texttt{constant} \tag{6.4}$$

where we used the fact that the variance of the principal components are the eigenvalues of the covariance matrix λ_1 and λ_2.
Assume that the covariance matrix is given by

$$COV[\vec{X}] = \begin{pmatrix} 4 & 2 \\ 2 & 4 \end{pmatrix} \tag{6.5}$$

Say we have an impact point $(6,6)$. Could that shell have originated from our gun with average impact point $(0,0)$? So, we test again $(E[X], E[Y]) = (0,0)$ against the alternative $(E[X], E[Y]) \neq (0,0)$. And we assume for both hypothesis, the covariance matrix the same and given in 6.5. The covariance matrix is also supposed to be known. The eigenvectors of our covariance matrix are $\vec{\nu}_1 = (1/\sqrt{2}, 1/\sqrt{2})$ and $\vec{\nu}_2 = (-1/\sqrt{2}, 1/\sqrt{2})$. The corresponding eigenvalues are 6 and 2. When we represent our impact point $(6,6)$ in the coordinate system of the principal components, we get: first coordinate is $12/\sqrt{2}$ and 0. Hence, the test statistic is

$$\frac{(\vec{X} \cdot \vec{\nu}_1)^2}{\lambda_1} + \frac{(\vec{X} \cdot \vec{\nu}_2)^2}{\lambda_2} = \frac{12^2}{2 \cdot 6} + \frac{0}{2} = 12$$

which is significant on the 5%-level. So, it probably was not our gun shooting this round. Now, note that in the coordinate system of the principal components, the covaraince matrix is diagonal since the components are independent and have covariance 0. So, the covariance matrix for the impact point expressed in the principal components is

$$COV\left[\left(\begin{array}{c} X_{PC} \\ Y_{PC} \end{array}\right)\right] = \left(\begin{array}{cc} \lambda_1 & 0 \\ 0 & \lambda_2 \end{array}\right) \tag{6.6}$$

So, our test statistic given on the left side of inequality 6.4 can be written as:

$$(X_{PC}, Y_{PC}) \cdot \left(COV\left[\left(\begin{array}{c} X_{PC} \\ Y_{PC} \end{array}\right)\right]\right)^{-1} \cdot \left(\begin{array}{c} X_{PC} \\ Y_{PC} \end{array}\right) \tag{6.7}$$

Now, if Σ denotes the covariance matrix, we have thus as acceptance region:

$$\vec{X} \cdot \Sigma^{-1} \vec{X} \leq \texttt{constant} \tag{6.8}$$

where as usual the constant depends on the significance level wanted. (To obtain 6.8 from 6.7, we used the fact that the expression 6.7 is invariant under an orthonormal basis change. And in our case, the coordinate system of the principal components are orthonormal. So, basically, the right side of 6.8 is just the right side of 6.7 rewritten in the original basis.

there is an additional justification for 6.8. Assume that we want to test our hypothesis of a normal impact point with $(E[X], E[Y]) = (0, 0)$ and given covariance Σ agains the alternative of a uniform random variable in an enormous erea around the origin denoted by R. We assume that the acceptance region has to be in R. So, this is a simple hypothesis testing situation. To find an optimal test, we simply, get the log ratio. We find the log ratio to be

$$\log \frac{f_{\texttt{normal},\mu_X=\mu_Y=0,\Sigma}}{f_{\texttt{uniform}}} = \texttt{constant} \cdot (-(\texttt{x},\texttt{y})\Sigma^{-1}\left(\begin{array}{c} x \\ y \end{array}\right))$$

Hence, putting the above expression below a constant leads to an optimal test when we test agains a uniform in a large area. So, we get again an acceptance region like this:

$$(x, y) \cdot \Sigma^{-1} \left(\begin{array}{c} x \\ y \end{array}\right) \leq \texttt{constant} \tag{6.9}$$

6.1 The mulitdimensional T-square test

Often you are in a situation that you do not have the covariance matrix but have to estimate it. Now, if you do not have a lot of data points this might introduce quite an error. More specifically, our experience with real high dimensional data is that we need a sample size about ten times bigger than the dimension of the random vector under consideration to have a good estimate of the covariance matrix. In that case, you can just act as if the true covariance was equal to the estimated one and replace in 6.9, the covariance by its estimate and still act as if 6.9 was Chi-square with p degrees of freedom.

Now, when your sample size n is less than about ten times the dimension p of your vector then you could have a subsancial error in your estimated covariance. So, in other words the difference between the true covariance and the estimated one, could be so big, that it is absolutely no longer justified to act as if they would be the same. In that case, however, we can show that the test statistic under H_0 does not depend on Σ or the expected value under H_0 and has an F distribution when properly rescaled with p and $n-p$ degrees of freedom.

Let us imagine a simple example: you would have the same number of male and female birdies in a sample. And you measure as usual length and wing span. Now, you want to find if there is a significant difference between male and female for the vector $\vec{X} = (X, Y)$. So, take the differences bird by bird. (This is only possible if you have the same sample size for female and male birdies). That is let (X_i, Y_i) be the difference between the i-th male and the ith female. (the ordering for each group is done at random). So you assume that you have p males and p females. Formally you want to test given the sample

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_p, Y_p)$$

the hypothesis H_0 that expectation is zero, that is the hypothesis that:

$$E[X] = E[Y] = 0.$$

Typically, you do not know the covariance matrix, so you have to estimate it. Also, as test statistic you take the average lengths:

$$\bar{X} = \frac{X_1, X_2, \ldots, X_p}{p} \; ; \; \bar{Y} = \frac{Y_1, Y_2, \ldots, Y_p}{p}$$

But, note:

$$COV(\bar{X}, \bar{Y}) = COV(\frac{X_1 + \ldots + X_n}{n}, \frac{Y_1 + \ldots + Y_n}{n}) = \frac{1}{n^2} COV(X_1 + \ldots + X_n, Y_1 + \ldots + Y_n) =$$
$$= \frac{1}{n^2} \sum_{i,j} COV(X_i, Y_j) = \frac{1}{n^2} \sum_{i,i} COV(X_i, Y_i) = \frac{n}{n^2} COV(X_1, Y_1) = \frac{COV(X_1, Y_1)}{n},$$

where we used that the covariance $COV(X_i, Y_j)$ is 0 when $i \neq j$. Similarly,

$$COV(\bar{X}, \bar{X}) = VAR(\bar{(}X)) = VAR[(X_1 + \ldots + X_n)/n] = VAR[X_1]/n = \frac{COV(X_1, X_1)}{n}$$

Samething for Y, we have $COV(\bar{Y}, \bar{Y}) = COV(Y, Y)/n$. **So, when instead of one birdy we consider their average, then the covariance matrix gets devided by n.** Hence, if we knew the covariance matrix, then the test statistics would be

$$(\bar{X}, \bar{Y}) \cdot \begin{pmatrix} COV(\bar{X}, \bar{X} & COV(\bar{X}, \bar{Y}) \\ COV(\bar{Y}, \bar{X}) & COV(\bar{Y}, \bar{Y}) \end{pmatrix}^{-1} \cdot \begin{pmatrix} \bar{X} \\ \bar{Y} \end{pmatrix} = n \cdot (\bar{X}, \bar{Y}) \cdot \begin{pmatrix} COV(X, X) & COV(X, Y) \\ COV(Y, X) & COV(Y, Y) \end{pmatrix}^{-1} \cdot \begin{pmatrix} \bar{X} \\ \bar{Y} \end{pmatrix} \tag{6.10}$$

But, the problem is we do not know the covariance matrix, so instead of the true covariance matrix, we will use the estimated one. So, our test statistics will be the expression on the right of 6.10. In other words, the acceptance region will be of the type:

$$n \cdot (\bar{X}, \bar{Y}) \cdot \left(C\hat{O}V \left[\begin{pmatrix} X \\ Y \end{pmatrix} \right] \right)^{-1} \cdot \begin{pmatrix} \bar{X} \\ \bar{Y} \end{pmatrix} \leq constant \tag{6.11}$$

where *constant* > 0 is a constant which depends on the significance level one choses. Let A be the true covariance matrix of the vector $\vec{Z} = (X, Y)$ to the power -0.5:

$$A := \left(COV \left[\begin{pmatrix} X \\ Y \end{pmatrix} \right] \right)^{-0.5}$$

For defining A, we leave the eigenvectors (principal components) identical and take the inverse square-roots of the eigenvalues. Then, $(X_i, Y_i)A$ is a normal vector with standard normal independent entries assuming that X_i and Y_i have expectation 0. To prove this note that the covariance matrix of a random row vector can be written as expectation of the vector mutliplied by its transpose. We have to multiply it from right otherwise we would get scalar product and not a matrix. So, in other words, the covariance matrix of $(X_i, Y_i)A$ is equal to

$$E[((X_i, Y_i)A)^t (X_i, Y_i)A] = E[A^t (X_i, Y_i)^t (X_i, Y_i)A] = A^t E[(X_i, Y_i)^t (X_i, Y_i)]A =$$

$$= A^t COV \left[\begin{pmatrix} X \\ Y \end{pmatrix} \right] A = Id$$

where Id is the 2×2 idendity matrix. We used, that since $A = COV[\vec{Z}]^{-0.5}$ then we have

$$A^t COV[\vec{Z}]A = Id.$$

Same thing is of course, $A(X, Y)^t$ has standard normal independent entries under H_0. Now, we simply apply this to our test statistic. We get that our test statistics 6.11 is equal to:

$$n \cdot (\bar{X}, \bar{Y})A \cdot \left(A \cdot C\hat{O}V \left[\begin{pmatrix} X \\ Y \end{pmatrix} \right] A \right)^{-1} \cdot A \begin{pmatrix} \bar{X} \\ \bar{Y} \end{pmatrix} \leq constant \tag{6.12}$$

Now having put the matrix A in the above test statistics we get that the test statistics is the test statistics for standartised data. That means if our original true covariance matrix would the identity, that is if X and Y would be standard normal and independent of each other and we would compute the statistics 6.11, then we would have the same as 6.12. For this let

$$(U_{X,i}, U_{Y,i}) := (X_i, Y_i)A \tag{6.13}$$

be the standardazied data. Then, by taking the transpose of 6.13, we also have:

$$\begin{pmatrix} U_{X,i} \\ U_{Y,i} \end{pmatrix} = A \begin{pmatrix} X_i \\ Y_i \end{pmatrix}. \tag{6.14}$$

Let

$$\bar{U}_X := \frac{U_{X,1} + \ldots + U_{X,n}}{n}$$

and let

$$\bar{U}_Y := \frac{U_{Y,1} + \ldots + U_{Y,n}}{n}.$$

By taking the average of 6.13 and 6.14, we find

$$(\bar{U}_X, \bar{U}_Y) := (\bar{X}, \bar{Y})A \quad , \quad \left(\begin{array}{c} \bar{U}_X \\ \bar{U}_Y \end{array} \right) = A \left(\begin{array}{c} \bar{X} \\ \bar{Y} \end{array} \right). \tag{6.15}$$

Then, we apply 6.13, 6.14 and 6.15 to 6.12, and find that 6.12 (and hence also 6.11) is equal to to:

$$n \cdot (\bar{U}_X, \bar{U}_Y) \left(C\hat{O}V \left[\left(\begin{array}{c} U_X \\ U_y \end{array} \right) \right] \right)^{-1} \cdot \left(\begin{array}{c} \bar{U}_X \\ \bar{U}_Y \end{array} \right) \leq constant \tag{6.16}$$

where

$$C\hat{O}V \left[\left(\begin{array}{c} U_X \\ U_Y \end{array} \right) \right] =$$

$$= \frac{1}{n-1} \left(\begin{array}{cccc} U_{X,1} - \bar{U}_X & U_{X,2} - \bar{U}_X & \ldots & U_{X,n} - \bar{U}_X \\ U_{Y,1} - \bar{U}_Y & U_{Y,2} - \bar{U}_Y & \ldots & U_{Y,n} - \bar{U}_Y \end{array} \right) \cdot \left(\begin{array}{cc} U_{X,1} - \bar{U}_X & U_{Y,1} - \bar{U}_Y \\ U_{X,2} - \bar{U}_X & U_{Y,2} - \bar{U}_Y \\ \multicolumn{2}{c}{\ldots} \\ U_{X,n} - \bar{U}_X & U_{Y,n} - \bar{U}_Y \end{array} \right)$$

designates the estimated covariance matrix, when we use the standartized data instead of the original one. The estimated covariance matrix for the standartized data replaces in formula 6.12 the expression

$$A \cdot C\hat{O}V \left[\left(\begin{array}{c} X \\ Y \end{array} \right) \right] A.$$

This is justified, because:

$$A \cdot C\hat{O}V \left[\begin{pmatrix} X \\ Y \end{pmatrix} \right] A =$$

$$A \cdot \frac{1}{n-1} \begin{pmatrix} X_1 - \bar{X} & X_2 - \bar{X} & \dots & X_n - \bar{X} \\ Y_1 - \bar{Y}_1 & Y_2 - \bar{Y} & \dots & Y_n - \bar{Y} \end{pmatrix} \cdot \begin{pmatrix} X_1 - \bar{X} & Y_1 - \bar{Y} \\ X_2 - \bar{X} & Y_2 - \bar{Y} \\ \dots & \\ X_n - \bar{X} & Y_n - \bar{Y} \end{pmatrix} \cdot A =$$

$$= \frac{1}{n-1} \begin{pmatrix} A \cdot (X_1 - \bar{X}, X_2 - \bar{X}, \dots, X_n - \bar{X}) \\ A \cdot (Y_1 - \bar{Y}\ , Y_2 - \bar{Y}\ , \dots, Y_n - \bar{Y}\ ) \end{pmatrix} \cdot \begin{pmatrix} (X_1 - \bar{X}, Y_1 - \bar{Y}) \cdot A \\ (X_2 - \bar{X}, Y_2 - \bar{Y}) \cdot A \\ \dots \\ (X_n - \bar{X}, Y_n - \bar{Y}) \cdot A \end{pmatrix} =$$

$$= \frac{1}{n-1} \begin{pmatrix} U_{X,1} - \bar{U}_X & U_{X,2} - \bar{U}_X & \dots & U_{X,n} - \bar{U}_X \\ U_{Y,1} - \bar{U}_Y & U_{Y,2} - \bar{U}_Y & \dots & U_{Y,n} - \bar{U}_Y \end{pmatrix} \cdot \begin{pmatrix} U_{X,1} - \bar{U}_X & U_{Y,1} - \bar{U}_Y \\ U_{X,2} - \bar{U}_X & U_{Y,2} - \bar{U}_Y \\ \dots & \\ U_{X,n} - \bar{U}_X & U_{Y,n} - \bar{U}_Y \end{pmatrix} =$$

$$= C\hat{O}V \left[\begin{pmatrix} U_X \\ U_Y \end{pmatrix} \right]$$

where the last equation above was obtained by using 6.13, 6.14 and 6.15. Now, we are going to devide the vector $(\bar{U}_X, \bar{U}_Y)$ by its Euclidian norm so as to get a unit vector:

$$(\bar{U}_X^*, \bar{U}_Y^*) = \frac{(\bar{U}_X, \bar{U}_Y)}{\sqrt{\bar{U}_X^2 + \bar{U}_Y^2}}.$$

with this notation, our inequality 6.16 reads:

$$n \cdot (\bar{U}_X^2 + \bar{U}_Y^2) \quad \cdot \quad (\bar{U}_X^*, \bar{U}_Y^*) \left(C\hat{O}V \left[\begin{pmatrix} U_X \\ U_y \end{pmatrix} \right] \right)^{-1} \begin{pmatrix} \bar{U}_X^* \\ \bar{U}_Y^* \end{pmatrix} \ \leq \ constant \tag{6.17}$$

Note that $\bar{U}_X$ and $\bar{U}_Y$ both are normal with expectation 0 and have standard deviation $1/\sqrt{n}$. Hence, the expression

$$n \cdot (\bar{U}_X^2 + \bar{U}_Y^2) \tag{6.18}$$

is the sum of two independent standard normals squared. Hence a Chi-square with 2 degrees of freedom. In general, if we have random vectors of dimension p, then this will be a Chi-square with p degrees of freedom. Now, note that

$$(U_{X,1} - \bar{U}_X, U_{X,2} - \bar{U}_X, \dots, U_{X,n} - \bar{U}_X) \tag{6.19}$$

is independent of $\bar{U}_X$. (To check just realize that $COV(U_{X,i} - \bar{U}_X, \bar{U}_X) = 0$ and that for joint normals we have that covariance 0 implies independence). Similarly, we get that

$$(U_{Y,1} - \bar{U}_Y, U_{Y,2} - \bar{U}_Y, \dots, U_{Y,n} - \bar{U}_Y) \tag{6.20}$$

is independent of $\bar{U}_Y$. Now the estimated covariance matrix for the standartize data only depends on 6.19 and 6.20 and hence is independent of $(\bar{X},\bar{Y})$. Also, by isotropy of space under a normal vector with independent standard normal entries, we find that $\bar{X}^2+\bar{Y}^2$ is independent of $(\bar{X}^*,\bar{Y}^*)$. This is to say, that in 6.17 the Chi-square variable give by $n(\bar{U}_X^2+\bar{U}_Y^2)$ is independent of the rest of expression 6.17. The rest of the expression is $(n-1)$ over a Chi-square with $n-p$ degrees of freedom. This then leads that our test statistics given in 6.11, is equal to:

$$T_0^2 := \frac{(n-1)p}{n-p}F_{p,n-p}$$

where $F_{p,n-p}$ denotes an F-statistics with p and $n-p$ degrees of freedom. Here p is the size of the vectors we consider, and n is the sample size. The only things which we still need to prove is that the other part (other than $n(\bar{U}_X^2+\bar{U}_Y^2)$) in 6.17, that is expression:

$$(\bar{U}_X^*,\bar{U}_Y^*)\left(C\hat{O}V\left[\left(\begin{array}{c} U_X \\ U_y \end{array}\right)\right]\right)^{-1}\left(\begin{array}{c} \bar{U}_X^* \\ \bar{U}_Y^* \end{array}\right) \tag{6.21}$$

is $(n-1)$ divided by a Chi-square matrix with $n-p$ degrees of freedom. This is done as follows:
define the $p\times n$ matrix L as follows:

$$L := \left(\begin{array}{cccc} X_1-\bar{X} & X_2-\bar{X} & \dots & X_n-\bar{X} \\ Y_1-\bar{Y}_1 & Y_2-\bar{Y} & \dots & Y_n-\bar{Y} \end{array}\right).$$

then let $\vec{L}_i$ be the i-th row of L. Hence, we have

$$L = \left(\begin{array}{c} \vec{L}_1 \\ \vec{L}_2 \end{array}\right)$$

Now let $P_1(.)$ denote the orthogonal projection onto the orthogonal complement of the span of $\vec{L}_2$ and the vectors $(1,1,\dots,1)^t$ in $\mathbb{R}^n$. (If there would be more than 2 coordinates X and Y, that is if there would be p coordinates, then we project orthogonally onto the orthogonal complement of the span of the vectors $\vec{L}_i$ with $i\neq 1$. Now, note that the vector $(\bar{(U)}_X^*,\bar{U}_Y^*)$ has a distribution which is invariant under rotation in space. Same thing for the estimated covariance matrix

$$C\hat{O}V\left[\begin{array}{c} U_X \\ U_Y \end{array}\right]$$

On top of this the two are independent. So, we can take any unit vector for $(\bar{U}_X^*,\bar{U}_Y^*)$ and we get the same distribution for 6.21. In other words, if we replace in expression 6.21, the unit vector $(\bar{U}_X^*,\bar{U}_Y^*)$ by $(1,0)$ we get the same distribution for 6.21. This allows us to realize that expression 6.21 has same distribution as

$$\begin{aligned}
&(1,0)\left(C\hat{O}V\left[\left(\begin{array}{c}U_X\\U_y\end{array}\right)\right]\right)^{-1}\left(\begin{array}{c}1\\0\end{array}\right)=\\
=&(n-1)\cdot(1,0)\left(L\cdot L^t\right)^{-1}\left(\begin{array}{c}1\\0\end{array}\right)=\\
=&(n-1)\cdot(1,0)\left((L^{-1})^t\cdot L^{-1}\right)\left(\begin{array}{c}1\\0\end{array}\right)=\\
=&(n-1)\cdot\left((1,0)(L^{-1})^t\right)\cdot\left(L^{-1}\left(\begin{array}{c}1\\0\end{array}\right)\right)=\\
=&(n-1)\cdot\left(L^{-1}\left(\begin{array}{c}1\\0\end{array}\right)\right)^t\cdot\left(L^{-1}\left(\begin{array}{c}1\\0\end{array}\right)\right)=\\
=&(n-1)\cdot\left|L^{-1}\left(\begin{array}{c}1\\0\end{array}\right)\right|_2^2=
\end{aligned}$$

where $|.|_2^2$ designates the Euclidian norm squared. Now the matrix L is not a square matrix, so what do we mean by L^{-1}? Well we restrict the map $\vec{x}\mapsto L\cdot\vec{x}$ from $\mathbb{R}^n$ to $\mathbb{R}^n$ to the imagespace of L. Now, by definition

$$P_1(\vec{L}_1)\cdot\vec{L}_2=0$$

and

$$P_1(\vec{L}_1)\cdot\vec{L}_1=|P_1(\vec{L}_1)|_2^2$$

hence, the vector

$$\frac{P_1(\vec{L}_1)}{|P_1(\vec{L}_1)|_2^2}$$

is transformed by the linear map L into the vector $(1,0)^t$. So, we can write

$$L^{-1}\left(\begin{array}{c}1\\0\end{array}\right)=\frac{P_1(\vec{L}_1)}{|P_1(\vec{L}_1)|_2^2}.$$

hence,

$$\left|L^{-1}\left(\begin{array}{c}1\\0\end{array}\right)\right|_2^2=\frac{P_1(\vec{L}_1)\cdot P_1(\vec{L}_1)}{|P_1(\vec{L}_1)|_2^4}=\frac{1}{|P_1(\vec{L}_1)|_2^2}$$

now $P_1(\vec{L}_1)$ is the orthogonal projection of a random vector with i.i.d. standard normal entries onto a $n-p$ dimensional space. hence, $|P_1(\vec{L}_1)|_2^2$ is a χ-square variable with $n-p$ degrees of freedom. This finishes our proof.

7 Asymptotics of testing, fisher information and the likelyhood ratio testing

So, assume that you have a bunch of variables which are i.i.d. $X_1, X_2, \ldots, X_n$ which have probability density $f(x, \theta_0)$, where the parameter θ_0 is not known. Assume, that we use maximum-likelyhood to estimate θ_0. How, precise will our estimate be in terms of having more date points n? We are going to show, that under some mild regularity conditions, when n is large enough, the maximum-Likelyhood estimate $\hat{\theta}_n$ is approximately normal with expectation θ_0 and variance given by

$$VAR[\hat{\theta}_n] \approx \frac{1}{n \cdot E[I(\theta_0)]}$$

where

$$I(\theta)$$

is the fisher information given by

$$I = -\frac{\partial^2 log(f(X, \theta))}{\partial^2 \theta}.$$

We call $l(x, \theta) := \log(f(x, \theta))$ the log likelyhood. The partial derivative according to θ is called the scoring function and denoted by S:

$$S = \frac{\partial l(x, \theta)}{\partial \theta} = \frac{\partial f(x, \theta)/\partial \theta}{f(x, \theta)}$$

So, we have

$$I(\theta) = -\frac{\partial S}{\partial \theta} = -\frac{\partial^2 f(x, \theta)/\partial^2 \theta}{f(x, \theta)} + \frac{(\partial f(x, \theta)/\partial \theta)^2}{f^2(x, \theta)}.$$

So, the maximum-likelyhood estimate is the value of θ which maximizes the log likelyhood. So, in other words

$$\hat{\theta}_n = argmax \sum_{i=1}^{n} l(X_i, \theta)$$

Now, the maximum is reached where the derivative is 0. So, we find that $\hat{\theta}_n$ is solution to

$$\frac{\sum_{i=1}^{n} S(X_i, \theta)}{n} = 0. \qquad (7.1)$$

The coefficient n does not change anything, we added it for later it will be easier to calculate. Now, Let us calculate the expected score at θ_0:

$$E[S(X, \theta_0)] =$$

$$= E[\frac{\partial f}{\partial \theta}(X, \theta_0) \frac{1}{f(x, \theta_0)}] = \int \frac{\partial f}{\partial \theta}(X, \theta_0) \frac{1}{f(x, \theta_0)} f(X, \theta_0) dx = \int \frac{\partial f}{\partial \theta}(X, \theta_0) dx = \frac{\partial \left(\int f(x, \theta) dx \right)}{\partial \theta} = 0$$

where we used that the integral of a probability density is always 1:

$$\int f(x,\theta)dx = 1, \forall\theta$$

and hence the derivative is 0. So,this shows that expression 7.1, has expectation 0 at θ_0. Now, the expression 7.1 being a sum of independents, it is approximately normal. Let us calculate the variance of $S(X,\theta_0)$:
We have

$$VAR[S(X,\theta_0)] = E[S^2(X,\theta_0)] - E^2[S(X,\theta_0)] = E[S^2(X,\theta_0)] =$$
$$= E[\left(\frac{\partial f}{\partial\theta}(X,\theta_0)\right)^2 \frac{1}{f^2(x,\theta_0)}]$$

Hence, the variance of expression 7.1, being equal to $VAR[S(X,\theta_0)]/n$ is of order $O(1/n)$. Furthermore, the derivative of the expression on the left side of 7.1 is given by:

$$\frac{\sum_{i=1}^n \partial S(X_i,\theta)/\partial\theta}{n} = \sum_{i=1}^n \frac{I(X_i,\theta)}{n} \approx E[I(X,\theta)] \tag{7.2}$$

where the last approximation above comes from the law of large numbers which says that an average of i.i.d. variables is about equal to their expectation when $n \to \infty$. Now the expected fisher information can be calculated as follows:

$$E[I(\theta)] = \tag{7.3}$$
$$E[-\frac{\partial^2 f(x,\theta)/\partial^2\theta}{f(x,\theta)}] + E[\frac{(\partial f(x,\theta)/\partial\theta)^2}{f^2(x,\theta)}] = E[\frac{(\partial f(x,\theta)/\partial\theta)^2}{f^2(x,\theta)}], \tag{7.4}$$
$$\tag{7.5}$$

where we used

$$E[\frac{\partial^2 f(x,\theta)/\partial^2\theta}{f(x,\theta_0)}] =$$
$$= \int \frac{\partial^2 f}{\partial^2\theta}(X,\theta_0)\frac{1}{f(x,\theta_0)} f(x,\theta_0)dx = \int \frac{\partial^2 f}{\partial^2\theta}(X,\theta_0)dx = \frac{\partial^2 \int f(X,\theta)}{\partial^2\theta} = 0$$

Hence, we find

$$E[I(X,\theta_0)] = VAR[S(X,\theta_0)], \tag{7.6}$$

So, say you have a differentiable function $\theta \mapsto g(\theta)$ which take a small value of order $O(1/n)$ at θ_0 and has a derivative at θ_0 bounded away from zero of order $O(1)$. Then, there will typically by a solution to $g(\theta) = 0$ in the vicinity of θ_0 given approximately by

$$\theta = -\frac{g(\theta_0)}{g'(\theta_0)}.$$

(To see this, just put the first order Taylor expension:

$$g(\theta) \approx g(\theta_0) + \Delta\theta g'(\theta_0)$$

equal to 0 and solve for $\Delta\theta$). So, when we take expression on the left side of 7.1 as $g(\theta)$, then we find that the solution to 7.1 is approximately equal to:

$$\hat{\theta} \approx \frac{g(\theta_0)}{g'(\theta)} = \frac{\frac{\sum_{i=1}^{n} S(X_i,\theta)}{n}}{\frac{\partial \sum_{i=1}^{n} S(X_i,\theta)/\partial\theta}{n}} \approx \frac{\frac{\sum_{i=1}^{n} S(X_i,\theta)}{n}}{E[I(X,\theta_0)]}$$

where we used 7.2. The expression on the very right side of the last equality above, has expectation 0 and variance equal to

$$\frac{VAR[S]}{n \cdot E^2[I(X,\theta_0)]} = \frac{1}{n \cdot I(X,\theta_0)},$$

where we used equation 7.6. We have just finished showning, that as n becomes bigger the maximu-likelyhood estimate $\hat{\theta}_n$ is approximately normal with expectation 0 and variance given by:

$$VAR[\hat{\theta}_n] \approx \frac{1}{n \cdot E[I(X,\theta_0)]}.$$

7.1 The most precise testing one could get

Say you have two type of particles and you measure their size. One has size 1 and the other has size 2, what ever the units are. You try based on the measurment of the size to determine which type of particle it is. Now, you know that the particle you measure is one of these two types. Assume that you get to measure one particle, but the measurment error is of order 20. (Standard deviation of measurment error is about 20). Then, you would be unable to say what type of particle it is based on your measurment. (Error of first type plus error of second type sum up to about 1). Then, you buy a better measurment tool. Now, you make measurments with a precision of 0.1. (Standard deviation of measurment error is about 0.1) Now you will be able to say with high probability which type of particle you have, because the error is much smaller than the difference in size, and by recognising the size you figure out the particle type. So, basically this is about testing where H_0: X is normal with expectation μ_1 and standard deviation σ against H_1 : X is normal with expectation μ_2 and same standard deviation σ. So, lousely speaking when $\mu_1 - \mu_2| = \sigma$ is the breaking point where we start being able to distinguish between the two hypothesis. When σ is quite a bit bigger than $|\mu - 1 - \mu_2|$ then we can not recognize which hypothesis holds. When σ is quite a bit smaller than $\mu_1 - \mu_2|$, then we can detect which hypothesis we are dealing with with high probability. So, now we are going to apply this idea when we test that a bunch of i.i.d. points $X_1, X_2, \ldots, X_n$ are generated by the probability density $g(.)$ or $f(.)$. (so, they are all either generated by $f(.)$ or all by $g(.)$). Then we know that the optimal test will be a ratio test, with acceptance region given by:

$$\frac{f(X_1) \cdot f(X_2) \ldots f(X_n)}{g(X_1) \cdot g(X_2) \ldots g(X_n)} \geq cst$$

where *cst* is a constant which will determine the significance level. We can also take the *log*-ratio, which would lead to the acceptance region being:

$$\frac{\log(\frac{f(X_1}{g(X_1)}) + \log(\frac{f(X_2)}{g(X_2)}) + \ldots + \log(\frac{f(X_n)}{g(X_n)})}{n} > cst. \tag{7.7}$$

Now, this is a sum of independents and hence asymptotically normal. It will have different expectation depending on wether the "true" underlying model is with $f(.)$ or with $g(.)$. The difference in expectation is

$$\Delta E = E_f[\log(\frac{f(X}{g(X)}] - E_g[\log(\frac{f(X}{g(X)}] = \int[\log(\frac{f(X}{g(X)}f(x)dx - \int \log(\frac{f(X}{g(X)})g(x)dx$$

And we need to compare this to the standard deviation

$$\sqrt{VAR[\log(\frac{f(X}{g(X})]}$$

which in many cases we consider does not dependen a lot on which model we use. So, when the standard deviation is quite a bit smaller than the difference in expectation we can tell which model we are dealing with whilst otherwise we can not.
We are going to apply this idea to a one paratmeter family $f(x,\theta)$. here $f(x,\theta 0$ is a one parameter family. Assume the date is generated by $f(x,\theta_0$. Now, take any non-random $\theta \neq \theta_0$. If we don't know, can we test with our n data points to find out which parameter is the true one? That is can we figure out if it is θ_0 or θ? (We assume that θ is a fixed given value which is already a little close to θ_1. We simply need to compare the expectation to the variance of the log ratio statistics.. So, when we take $f(x) = f(x,\theta)$ and $g(x) = f(x,\theta_0)$, then our test-statistics 7.7 is:

$$\sum_{i=1}^{n} \log \frac{f(X_i,\theta)}{f(X_i,\theta_0)} = \sum_{i=1}^{n}(\log(f(X_i,\theta) - \log(f(X_i,\theta_0)) \approx \Delta\theta \sum_{i=1}^{n} \frac{\partial \log(f)}{\partial\theta}(X_i,\theta_0) = \Delta\theta \sum_{i=1}^{n} \frac{\frac{\partial f}{\partial \theta}}{f}(X_i,\theta_0).$$

Hence, the difference in expectation is going to be obtained by taking the expectation of the log ratio under $f(x,\theta)$ vs under $f(x,\theta_0)$. This gives for the difference in expectation of the log-ratio under the two different densities:

$$\begin{aligned}
\Delta E &= \Delta\theta \left(E_\theta[\sum_{i=1}^{n} \frac{\frac{\partial f}{\partial\theta}}{f}(X_i,\theta_0)] - E_{\theta_0}[\sum_{i=1}^{n} \frac{\frac{\partial f}{\partial\theta}}{f}(X_i,\theta_0)] \right) = \\
&= \Delta\theta \left(\int \sum_{i=1}^{n} \frac{\frac{\partial f}{\partial\theta}}{f}(X_i,\theta_0) f(x,\theta)dx - \int \sum_{i=1}^{n} \frac{\frac{\partial f}{\partial\theta}}{f}(X_i,\theta_0) f(x,\theta_0)dx \right) = \\
&= \Delta\theta \int \sum_{i=1}^{n} \frac{\frac{\partial f}{\partial\theta}}{f}(X_i,\theta_0)(f(x,\theta) - f(x,\theta_0)dx \approx \\
&\approx \Delta^2\theta \int \sum_{i=1}^{n} \frac{(\frac{\partial f}{\partial\theta})^2}{f}(X_i,\theta_0)dx = \\
&= n\Delta^2\theta \int \frac{(\frac{\partial f}{\partial\theta})^2}{f}(X_i,\theta_0)dx
\end{aligned}$$

the other thing we need to find is the variance of the log-likelyhood ratio. So, we have

$$\begin{aligned}
& VAR[\sum_{i=1}^{n} \log(\frac{f(X_i,\theta)}{f(X_i,\theta_0)})] = \sum_{i}^{n} VAR[\log(\frac{f(X_i,\theta)}{f(X_i,\theta_0)})] = \\
&= nVAR[\log(\frac{f(X,\theta)}{f(X,\theta_0)})] \approx nVAR[\Delta\theta \frac{\frac{\partial f}{\partial\theta}}{f}(X_i,\theta_0)] = \\
&= n\Delta^2\theta E\left[\left(\frac{\frac{\partial f}{\partial\theta}}{f}(X_i,\theta_0)\right)^2\right] = \\
&= n\Delta^2\theta \int \left(\frac{\frac{\partial f}{\partial\theta}}{f}(X_i,\theta_0)\right)^2 f(x,\theta_0)dx = \\
&= n\Delta^2\theta \int \frac{\left(\frac{\partial f}{\partial\theta}\right)^2}{f}(X_i,\theta_0)dx
\end{aligned}$$

Now, puting the difference in expectation equal to the variance of the log-likelyhood ratio will tell us at what point θ is too close to θ_0 to distinguish. So, we put

$$\Delta E = \sqrt{VAR}$$

which leads to

$$n\Delta^2\theta \int \frac{(\frac{\partial f}{\partial\theta})^2}{f}(X_i,\theta_0)dx = \sqrt{n\Delta^2\theta \int \frac{\left(\frac{\partial f}{\partial\theta}\right)^2}{f}(X_i,\theta_0)dx}$$

which implies

$$\Delta\theta = \frac{1}{\sqrt{n \int \frac{\left(\frac{\partial f}{\partial\theta}\right)^2}{f}(X_i,\theta_0)dx}} = \frac{1}{\sqrt{n \cdot I(X)}}$$

where we used

$$E[I(X,\theta)] = \int \frac{\frac{\partial f}{\partial \theta}}{f} dx. \qquad (7.8)$$

which was proven in the previous section in (see equation 7.4).

7.2 The likelihood-ratio testing and generalised linear models