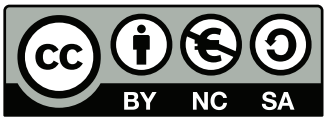
Métodos Avanzados en Estadística

Estudiante

Máster en Ciencia de Datos

Universidad Autónoma de Madrid

tu-web.es



Métodos Avanzados en Estadística

Estudiante

Máster en Ciencia de Datos

Universidad Autónoma de Madrid

tu-web.es

Índice

I. Bootstrap **5**

II. PDF Estimation **6**

1. PDF Estimation **6**
1.1. Estimating the kernel as a convolution 6
1.2. Integrating the MSE 7
1.2.1. Other criteria 7

2. Bias and Variance approximations **8**
2.0.1. Optimal kernel 9
2.0.2. Choosing the smoothing parametern 10
2.1. Multivarian density estimation 11
2.1.1. The curse of dimensionality 11

III. Regression **12**

3. Settlement **12**

4. Non parametric regression. Nadaraya-Watson Estimator **13**
4.1. Integrated MSE of the N-W estimator 14
4.2. Penalized MSE 15
4.3. Multiple linear regression 15
4.4. MSE fit 16
4.4.1. Variability decomposition 17
4.5. Reduced and complete models 18
4.5.1. Prediction errors 18

IV. Classification **21**

Parte I.
Bootstrap

Parte II.
PDF Estimation

1. PDF Estimation

Let $X_1, \ldots, X_n$ be an iid sample from the distribution F with p.d.f. f. We woud like to estimate f making no previous assumptions, just making use of the data. Our goal will be to compute an estimator $\hat{f}$ such that $\hat{f} \approx f$. Let $h \approx 0$, then

$$P(x-h \leq X \leq x+h) = \int_{x-h}^{x+h} f(t)dt \approx 2hf(x)$$

that means that we can approximate $f(x)$ as follows:

$$f(x) \approx \frac{1}{2h} P(x-h \leq X \leq x+h).$$

If we replace the probability by the proportion we obtain an p.d.f. estimator:

$$\hat{f}(x) = \frac{1}{2h} \frac{\#\{i : |x-X_i| < h\}}{n} = \frac{1}{2hn} \#\{i : \frac{|x-X_i|}{h} \leq 1\}.$$

If $K(x) = \frac{1}{2}\mathbb{I}_{\{|x|\leq 1\}}$, where $\mathbb{I}_A$ is the indicator function over A, then

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)$$

1.1. Estimating the kernel as a convolution

Let $U = Y + Z$ where:

1. $Y \sim F_n$ where F_n is the empiric distribution function
2. Z's density function is $K_h(x) = h^{-1}K(x/h)$.
3. Y and Z are independent.

that is, making a small perturbation on the empiric distribution. We can understand the kernel estimator as a slightly modified(smoothed) empiric distribution so that it is continuous.

Proposition 1.1. In the previous conditions, the p.d.f. of U is the kernel estimator $\hat{f}$.

Demostración. Firstly, we recall that

$$F_u(x) = P(u \leq x) = P(y+z \leq x)$$

now, using the total probability formula, we obtain

$$P(y+z\leq x)=\frac{1}{n}\sum_{i=1}^{n}P(y+z\leq x|y=x_i)=\frac{1}{n}\sum_{i=1}^{n}P(z\leq x-x_i)$$

in the last step, we can *forget* the conditional since Z, Y are independent. Then, we can use the definition of the probability to obtain:

$$\frac{1}{n}\sum_{i=1}^{n}P(z\leq x-x_i)=\frac{1}{nh}\sum_{i=1}^{n}\int_{-\infty}^{x-x_i}k(\frac{u}{h})du$$

We can now derivate the distribution function to obtain

$$f_u(x)=F_u'(x)=\frac{1}{nh}=\sum_{i=1}^{n}k(\frac{x-x_i}{h})=\hat{f}(x)$$

□

Using this, we can have an *algorithm* to obtain an iid random variable with p.d.f. $\hat{f}$ following the next steps:

1. Pick a random element with the same probability in the set $X_1,\ldots,X_n$. Let the result be X^*.
2. Simulate items to obtain Z with distribution K_h
3. Compute $U=X^*+Z$

1.2. Integrating the MSE

In general, we use the bias and the variance of an estimator to determine its *goodness*. Recall that

1. The bias of $\hat{f}(x)$ is $E[\hat{f}(x)]-f(x)$
2. The variance of $\hat{f}(x)$ is $E[(\hat{f}(x)-E[\hat{f}(x)])^2]$
3. The MSE $E[\hat{f}(x)-f(x)]^2$ verifies:

$$MSE=Bias^2[\hat{f}(x)]+Var[\hat{f}(x)].$$

We would like to determine the MSE in terms of the hyperparameters h, n and the smoothness of f. If we use higer values for h, then we would have a high bias since we would be forgetting the data. If we use very small values for h, our estimator would be very colse to the data, we would have a high variance and very small bias.

1.2.1. Other criteria

There are also other criteria to integrate the MSE. For instance, considering the $||\dot{|}|_p$ in L_p, we can see that the ECMI is the expectation of the L_2 distance between f and $\hat{f}$ squared, that is:

$$ECMI(\hat{f})=E[||f-\hat{f}||_2^2]$$

It is easily shown seeing that:

$$E(||f-\hat{f}||_2^2) = E[\int |f(x)-\hat{f}(x)|^2 dx]$$

and, using Fubini's theorem, we obtain:

$$E[\int |f(x)-\hat{f}(x)|^2 dx] = \int E[(f(x)-\hat{f}(x))^2]dx = \int MSE(x)dx$$

2. Bias and Variance approximations

We will make the following assumptions for this section:

- K is a symmetric function with $\int K(u)du = 1$, $\int uK(u)du = 0$.
- $\sigma_K^2 = \int u^2 K(u)du < \infty$ and $d_k = ||K||_2^2 = \int K(u)^2 du < \infty$.
- f is twice *derivable* with continuous derivative.

Proposition 2.1. The bias... (diapositiva 21/37)

Demostración. Then, the approximation of the bias can be obtained as follows:

$$E[\hat{f}(x)] = E\left[\frac{1}{nh} = \sum K\left(\frac{x-x_i}{h}\right)\right] = \frac{1}{n}E[K\left(\frac{x-x_i}{h}\right] = \frac{1}{n}\int K\left(\frac{x-t}{h}f(t)dt\right)$$

we apply a *variable change* $w = \frac{x-t}{h}$, $dw = \frac{-dt}{h}$, $t = x-wh$, we obtain

$$\frac{1}{n}\int K\left(\frac{x-t}{h}f(t)dt\right) = \int K(w)f(x-wh)dw$$

and, applying Taylor's theorem (forgetting the terms that have h^3 and so on since h will be very small),

$$\int K(w)f(x-wh)dw = \int K(w)\left[f(x)-whf'(x)+\frac{w^2h^2}{2}f''(x)'\right]dw = f(x)+\frac{h^2}{2}f''(x)\int w^2K(w)dw$$

where the last integral equals σ_k^2, so

$$Bias(\hat{f}(x)) = \frac{h^2}{2}f''(x)\sigma_k^2.$$

As we can see, the bias of $\hat{f}$ depends highly on h.

□

Proposition 2.2. The variance..

Demostración. We can see that

$$Var[\hat{f}(x)] = Var\left[\frac{1}{nh} = \sum K\left(\frac{x-x_i}{h}\right)\right] = \frac{1}{h^2 n} Var\left[K\left(K(\frac{x-x_i}{h})\right)\right] = \frac{1}{h^2 n}\left[E\left[K^2(\frac{x-x_i}{h})\right] - E^2\left[K(\frac{x-}{h}\right.\right.$$

which is approximately

$$\frac{1}{h^2 n} E\left[K^2(\frac{x-x_i}{h})\right] = \frac{1}{nh^2}\int K^2(\frac{x-t}{h}) f(t) dt$$

and, applying the same *variable change*, and Taylor's theorem:

$$= \frac{1}{nh^2}\int K^2(w) f(x-hw) dw \approx \frac{1}{nh}\int K^2(w)\left[f(x) - hwf'(x) + \frac{h^2 w^2}{2} f''(x)\right] dw$$

Since we are dividing by nh, we can forget the term $f'(x)$ term, so we finally obtain:

$$Var[\hat{f}] = \frac{1}{nh} f(x) d_k$$

□

Using both previous propositions, we obtain:

Proposition 2.3. The approximation of the MSEI is:

$$MSEI(\hat{f}) \approx \frac{h^4}{4}\sigma_k^4 \int f''(x)^2 dx + \frac{d_K}{nh} = \frac{h^4}{4}\sigma_K ||f''||_2^2 + \frac{||K||_2^2}{nh}$$

The idea now is to optimize this function with respect h. We have to derivate this function respect to h and make it equal to 0, that is:

$$\frac{4h^3}{4}\sigma_K^4 ||f''||_2^2 = \frac{||K||_2^2}{nh^2}$$

and the optimal value is

$$h^* = cn^{-1/5} \to 0$$

which implies

$$nh^* = cn^{4/5} \to \infty$$

The conclusions are:

- h^* has to go to zero but slowly
- The speed at which h goes to zero in this case is $n^{-4/5}$, which is pretty slower than just n.

2.0.1. Optimal kernel

What is more, when computing the IMSE, the factor that depends on K is $\sigma(K) = \sigma_K^{4/5} ||K||_2^{8/5}$. It can be proved that if $K_h(\cdot) = \frac{1}{h} K(\frac{\cdot}{h})$.

To optimize the kernel, we have the following variational problem:

$$\min ||K||_2^2 \quad s.t.$$

The solution to this problem is the Epanechnikov

$$K^*(u) = \frac{3}{4\sqrt{5}}\left(1 - \frac{t^2}{5}\right) \quad \text{if } -\sqrt{5} \leq u \leq \sqrt{5}$$

The choice of the kernel does not hardly affect the value of the IMSE.

2.0.2. Choosing the smoothing parametern

To select the optimal smoothing parameter there are many methods:

1. Plug in:
   - Suppose that $f \approx N(\mu, \sigma)$
   - Non parametric methods
2. Cross validation

Plug-in supposing Normal.-
We assume that

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp ...$$

So, we substitute $\|f''\|_2^2$ in the expression of the optimal parameter h^* and , since we already know that the value of the norm, we obtain

$$h^* = \sigma(4/(3n))^{1/5} \approx \sigma(1.0456)n^{-1/5}$$

So, since the number is approximately 1, it is enough to estimate σ to estimate h^*. Some literature (*Silverman*) proposes to use $\hat{\sigma} = \min\{s, \hat{\sigma}_{ri}\}$ where $\hat{\sigma}_{ri}$ is the interquartilic range (estandardized so that it converges to σ).

Non-parametric plug-in.-
We fix a preliminar smoothing parameter g, using for instance the previous approximation. This way, we obtain an auxiliar estimator $\hat{f}_g(x)$. Using this estimator, we now estimate

$$\|\hat{f''}\|_2^2 = \int \hat{f}_g''(x)^2 dx$$

Cross Validation
The intuitive idea under cross validation is to divide the population in two splits and we use one of them to obtain information of how good is the estimator computed with the first one. We can then compute

$$MSEI(\hat{f}) = E\left[\int \hat{f}(x;h)^2 dx\right] - 2E\left[\int \hat{f}(x;h) f(x) dx\right] + \int f(x)^2 dx$$

Again, we would like to find h^* such that this IMSE is small. As we can see, the last term only depends on the *true* density function. For each h, our sample is fixed, so the expectation is reduced to a single term and we would like to minimize

$$C(h) = \int \hat{f}(x;h)^2 dx - \frac{2}{n}\sum \hat{f}_{(-i)}(X_i;h)$$

where $f_{(-i)}$ is the estimator computed with all the observations except X_i. The last term comes from the following reasoning. Let $x_1, \ldots, x_n \to \hat{f}$, and let us have another point x_{n+1}. Then, we can see that:

$$\int \hat{f}(x;h) f(x) dx = E\left[\hat{f}(x_{n+1};h) | x_1, \ldots x_n\right]$$

If we take expectation in both sides and use the known formula, we obtain:

$$E\left[\int \hat{f}(x;h) f(x) dx\right] = E\left[\hat{f}(x_{n+1};h)\right]$$

2.1. Multivarian density estimation

Consider d dimensions, we have that

$$\hat{f}(x) = \frac{1}{n|H|}\sum \tilde{K}(H^{-1}(x - X_i)), \quad x = (x_1, \ldots, x_d) \in \mathbb{R}^d$$

where $\tilde{K}$ is a multivariant density, $\Sigma \in \mathscr{M}_{d\times d}$ positive definited (?), $H = \Sigma^{1/2}$ and $|H|$ is the determinant of H. Note that, if $\Sigma = CDC'$ where D is a diagonal matrix, we define $\Sigma^{1/2} = CD^{1/2}C'$ where $D^{1/2}$ has the squared roots of the elements. The problem of this definition is that there is a wide range of parameters that we have to fix.

We approach this problem by simplyfying the expression:

- We consider $\tilde{K}$ as the product of identic unidimensional kernels:

$$\tilde{K}(x_1, \ldots, x_d) = K(x_1)\ldots K(x_d)$$

- We consider H to be diagonal and each element in H to be the same, that is: $H = hI_d$.

With this considerations, we obtain

$$\hat{f}(x) = \frac{1}{nh^d}\sum_{i=1}^{n}\prod_{j=1}^{d} K\left(\frac{x_j - X_{i,j}}{h}\right), \quad x = (x_1, \ldots, x_d) \in \mathbb{R}^d$$

2.1.1. The curse of dimensionality

Unless we have huge amounts of data, it is quite hard to find data in many zones of the sample space, so the properties of our estimators are worsen.

For instance, it can be shown that in dimension d,

$$ECMI^* \approx O(n^{\frac{-4}{4+d}})$$

Parte III.
Regression

3. Settlement

The general goal in a regression problem is, given a random variable Y and a vector of random regressor variables $X = (X_1, \ldots, X_p)$, study the relation between X and Y. Remark that Y can also be a random vector and in this case we are in the case of multivariant regression.

Definition 3.1. In the previous conditions, we call the regression function to:

$$m(X) = E[Y|X], \quad m(x) = E[Y|X = x].$$

It is known that $m(X)$ is the best prediction of Y from X (using the MSE). Our goal is to estimate $\hat{m}(x)$ using a finite number of i.i.d. datapoints $(x_1, y_1), \ldots, (x_n, y_n)$ where $x_i \in \mathbb{R}^p$.

Usually, regression models are expressed as follows:

$$Y = m(X) + \epsilon,$$

where $E(\epsilon|X) = 0$ or, equivalently, $m(X) = E[Y|X]$. That can be proved the following way:

$$E[Y|X] = E[m(x) + \epsilon|x] = m(x) + E[\epsilon|x] = m(x)$$

because the expectation $E[\epsilon|x] = 0$. These hypothesis provide the following statements:

- $E(\epsilon) = E(E(\epsilon|X)) = E(0) = 0$
- $Var(\epsilon = 0)$ using the formula of the total variance.
- $E(\epsilon X) = E[E(\epsilon X|X)] = E[E(\epsilon|X)] = 0$.
- Lastly, the last item implies that $Cov(\epsilon, X) = E(\epsilon X) - E(\epsilon)E(X) = 0 - 0 \cdot E(X)$.

There are many different possible models:

1. Linear regression, where
$$m(X) = m(X_1, \ldots, X_p) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$
2. Non parametric models, that assume continuity or smoothness conditions for $m(x)$.

4. Non parametric regression. Nadaraya-Watson Estimator

Let us assume that the vector X, Y has a joint p.d.f. $f(x,y)$ and the marginal p.d.f. of X is $g(x)$. Then, using the definition of the conditional probability:

$$m(x) = E[Y|X=x] = \int y f(y|x) dy = \frac{\int y f(x,y) dy}{g(x)}.$$

A very good first idea is to use a kernel estimator in the numerator and in the denominator to replace the densities that appear in the last expression.

$$\hat{g}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

and

$$\hat{f}(x,y) = \frac{1}{nh^2} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) K\left(\frac{y - Y_i}{h}\right).$$

Proposition 4.1. In the previous conditions, the Nadaraya-Watson estimator is has the expression:

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) Y_i}{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)}$$

Demostración.

$$\int y \hat{f}(x,y) dy = \frac{1}{nh^2} \sum K\left(\frac{x - X_i}{h}\right) \int y K\left(\frac{x - X_i}{h}\right) dy$$
$$\overset{(1)}{=} \frac{1}{nh} \sum K\left(\frac{x - X_i}{h}\right) \underbrace{\int (uh + Y_i) K(u)\, du}_{(2)}$$

where, in (1) we have applied the change of variables $u = \frac{y - y_i}{h}$ and in (2) we have used that $\int u k(u) du = 0$ and that $\int k(u) du = 1$. To end this proof, we insert this term in the expression of $\hat{m}(x)$. □

Proposition 4.2. The N-W estimator in x is the value $\hat{\beta}_0$ that minimizes

$$\sum_{i=1}^{n} w_i(x)(Y_i - \beta_0)^2,$$

which is a sum of weighted least squares.

We can change the way of expressing $\hat{m}(x)$. If we name

$$w_i(x) = \frac{K\left(\frac{x-X_i}{h}\right)}{\sum K\left(\frac{x-X_i}{h}\right)},$$

then $\hat{m}(x) = \sum w_i(x) y_i$.

Remark 4.1. The estimator is an average of Y_i locally weighted such that, to estimate $m(x)$, we use only the Y_i such that $X_i \approx x$.

A natural question is to consider the extreme cases of the smoothing parameter h of the kernel.

- If $h \to \infty$, we have that

$$\lim_{h\to\infty} w_i(x) = \frac{K(0)}{nK(0)}\frac{1}{n}$$

and, then,

$$\hat{m}(x) = \sum \frac{1}{n} Y_i = \hat{Y}$$

- If $h \to 0$, since $K\left(\frac{x-X_i}{h}\right) \to 0$ when $j \neq i$ and $K(0)$ if $i = j$, we obtain

$$\lim_{h\to\infty} \hat{m}(x_i) = Y_i$$

We can generalize the previous case as follows:

Proposition 4.3. The N-W estimator can be expressed as

$$\hat{m}(x) = \hat{\beta}_0 + \hat{\beta}_1 x + \cdots + \hat{\beta}_p x^p$$

where $\hat{\beta}_1, \ldots, \hat{\beta}_p$ minimize the term:

$$\sum_{i=1}^{n} w_i(x)(\mathbf{Y}_i - \beta_0 - \beta_1(x_i - x) - \cdots - \beta_p(x_i - x)^p)^2.$$

4.1. Integrated MSE of the N-W estimator

We can also approximate the bias and the variance from the NW estimator.

Proposition 4.4. We can approximate the integrated variance of the N-W estimator as

$$\int Var[\hat{m}(x)]dx \approx \frac{\sigma^2 ||K||_2^2}{nh} \int \frac{dx}{h(x)}, \quad nh >>> 0.$$

The bias of this estimator can also be approximated as:

$$\int (E(\hat{m}(x) - m(x))^2)dx \approx \frac{h^4}{4}\sigma_K^4 \int \left(m''(x) + 2\frac{m'(x)g'(x)}{g(x)}\right)^2 dx$$

Note that:

- Since $g(X)$ is the density of X, if $g(x) \approx 0$, we do not have much information of X around x, so $\hat{m}(x)$ has a high variance.
- In the bias, the term $\frac{2m'(x)g'(x)}{g(x)}$ is a design bias depending on $\mathbf{X}$. This term disappears when locally linear regression is used.

4.2. Penalized MSE

In this case, the estimators are functions that minimize $\phi(\lambda)$ for $\lambda > 0$ and

$$\phi(\lambda) := \sum_i (Y_i - m(x_i))^2 + \lambda \int_a^b m''(x)^2 dx$$

The first term measures how our model fits the data. The second term controls the smoothness of our estimator. We can think of it as the regularization term used frequently in machine learning.

We have to consider the following extreme situations:

1. $$\lim_{\lambda \to \infty} \hat{m}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$
2. $$\lim_{\lambda \to 0} \hat{m}(x) = interpolator$$

Remark 4.2. The solution of this problem, which is a tradeoff between fitting our data and being smooth, is a spline.

Proposition 4.5. The function that minimizes $\phi(\lambda)$ is a cubic natural spline which nodes correspond to $x_1, \ldots, x_n$.

4.3. Multiple linear regression

The most common regression model is:

$$Y = \beta_0 + \sum_j^p \beta_j X_j + \epsilon$$

where $E(\epsilon | X_1, \ldots, X_p) = 0$ and its variance is σ^2. We can assume that, for many inferences (intervals, contrasts,...) $\epsilon | (X_1, \ldots, X_p)$ follows a gaussian distribution.

Each observation from the training sample follows the model:

$$y_i = \beta_0 + \sum_j^p \beta_j x_{i,j} + \epsilon_i, \quad i = 1, \ldots, n$$

where, again, $E(\epsilon_i | x_i) = 0$ and $Var(\epsilon_i | x_i) = \sigma^2$.

We can formulate the previous model as follows:

$$\underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}}_{\mathbf{Y}} = \underbrace{\begin{pmatrix} 1 & x_{1,1} & \ldots & x_{1,p} \\ 1 & x_{2,1} & \ldots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \ldots & x_{n,p} \end{pmatrix}}_{\mathbf{X}} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}}_{\beta} + \underbrace{\begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}}_{\epsilon}.$$

Using the names in the under braces, our problem is formulated as:

$$Y = X\beta + \epsilon, \quad \epsilon|X \equiv \mathrm{N}_n(0, \sigma^2 \mathbb{I}_n) \Leftrightarrow Y|X \equiv \mathrm{N}_n(X\beta, \sigma^2 \mathbb{I}_n)$$

we call X the design matrix.

Note that, throughout this section, the data X is fixed. Hence, in the expression $\mathbf{Y} = \mathbf{X}\beta + \epsilon$, the term $\mathbf{X}\beta$ is *constant*, so the distribution that $\mathbf{Y}$ follows depends only on the distribution of ϵ. Applying how the multi dimensional gaussian distribution is affected by sums and product by scalars, is how we obtain the distribution of $\mathbf{Y}|\mathbf{X}$.

4.4. MSE fit

This is the case where we want to fit our coefficients $\beta_0, \ldots, \beta_p$ using the MSE. We want to find the values $\beta_0, \ldots, \beta_p$ that minimize

$$\|\mathbf{Y} - \mathbf{X}\beta\|_2^2 = \sum_{i=1}^{n} \left[\mathbf{Y_i} - (\beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p})\right]^2$$

It can be shown that $\hat{\mathbf{Y}}$ is the orthogonal projection.

Definition 4.1. In the previous conditions, the residuals vector $\mathbf{e}$ is defined as

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

This residuals must be orthogonal to the columns of the design matrix.

$$\mathbf{X}'(\mathbf{Y} - \hat{\mathbf{Y}}) = 0 \iff \mathbf{X}'\mathbf{e} = 0$$

Proposition 4.6. The least squares estimators $\hat{\beta}$ is expressed as

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

To proof this, derive the error expression and then set it equal to zero to find the minimum.

Proposition 4.7. The least squares estimator is the maximum likelihood estimator of β.

$$L(\beta, \sigma^2) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left\{-\frac{1}{2\sigma^2 \|\mathbf{Y} - \mathbf{X}\beta\|_2^2}\right\}$$

Proposition 4.8. The vector $\hat{\beta}$ follows a $p+1$ gaussian distribution with means vector β and covariance matrix $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

$$\hat{\beta} \equiv N_{p+1}(\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}).$$

If we *fit the values*, we obtain the adjusted values vector, defined as:

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \mathbf{H}\mathbf{Y}, \quad \text{where } \mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'.$$

H is called the *hat matrix* and, geometrically, it is a projection matrix over V. Using this hat matrix, the residual vector is defined as:

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$$

We also want to estimate the **variance** σ^2. To do this, we can use the variance of our residuals:

$$\mathbf{S}_R^2 = \frac{1}{n-p-1}\sum_{i=1}^{n}\mathbf{e}_i^2$$

Proposition 4.9.

$$(n-p-1)\mathbf{S}_R^2/\sigma^2 \equiv \chi^2_{n-p-1}$$

This result is a consequence of the fact that the distribution of the residuals follows a gaussian distribution, and we are multiplying by a constant that normalizes each term, so we are adding Gaussian distributions, which is the definition of a χ^2

Note. Thanks to the previous proposition, we can build confidence intervals and contrasts for σ^2

Proposition 4.10. $\mathbf{S}_R^2$ and $\hat{\beta}$ are independent.

4.4.1. Variability decomposition

Definition 4.2. We can define the following quantities:

- The **Total squared sum**

$$SCT = \sum_{i=1}^{n}(\mathbf{Y}_i - \bar{\mathbf{Y}})^2$$

  measures the total variability of the predicted variable.
- The **Explained squared sum**

$$SCE = \sum_{i=1}^{n}(\hat{\mathbf{Y}}_i - \bar{\mathbf{Y}})^2$$

  measures the variability of the predicted variable *explained by the model*.
- The **Residual squared sum**

$$SCR = \sum_{i=1}^{n} e_i^2$$

  measures the variability of the predicted variable that is *not explained by the model*.

Using that the residuals are orthogonal to the regression variables, we obtain

Corollary 4.1. The total variability of the predicted variable can be decomposed into the the explained variability by the model and the residual variability (not explained by the model). That is:

$$SCT = SCE + SCR$$

Definition 4.3. The determination coefficient measures the capability of our model to explain $\mathbf{Y}$.

$$R^2 = \frac{SCE}{SCT}$$

Let us set in the case that we would like to contrast the hypothesis

$$\mathbf{H}_0 : \beta_1 = \cdots = \beta_p = 0.$$

In this case, we would use the statistic:

$$F = \frac{SCE/p}{SCR/(n-p-1)}.$$

Under $\mathbf{H}_0$, the statistic F follows a distribution $F_{p,n-p-1}$.

4.5. Reduced and complete models

Adding complexity to a model usually turns into a better fit to available data. However, this could result in worse predictions for new data. The simpler the model, the less overfitting.

Our goal in this section will be to compare two models:

1. The complete model
2. A simplified model **M0**, where we constrain our model to $\mathbf{A}\beta = 0$, where $\mathbf{A} \in \mathscr{M}_{k\times(p+1)}$ and $range(\mathbf{A} = k < p+1$.

We will compare both models contrasting $H_0 : \mathbf{A}\beta = 0$.

4.5.1. Prediction errors

Suppose that we have a real model, predicting some data. Let $Y = (Y_1, \ldots, Y_n)$ be the labels of the training data, $\tilde{Y} = (\tilde{Y}_1, \ldots, \tilde{Y}_n)$ the labels of the test data. We hypothesize that the vectors $Y, \tilde{Y}$ are independent. We assume that $E(Y) = E(\tilde{Y}) = \mu = (\mu_1, \ldots, \mu_n)$ and Σ be the covariance matrix.

We fit a regression model $Y = X\beta + \epsilon \iff Y = \mu + \epsilon$ where $\mu \in V = \{X\beta \ : \beta \in \mathbb{R}^p\}$. We already know that:

- The MSE model is $\hat{\beta} = (X'X)^{-1}X'Y$
- ... D(44/57)

Definition 4.4. We call the training error:

$$E_{\text{train}} E\left[\sum_{i=1}^{n}(Y_i - x_i'\hat{\beta})^2\right]$$

and we can define the test error the same way using test labels.

Proposition 4.11. The test error can be expressed as:

$$E_{\text{test}} = E_{\text{train}} + 2E(\|C\|^2)$$

Demostración. Recall that:

$$E_{\text{test}} = E(\|A\|^2) + E(\|B\|^2) + E(\|C\|^2) + 2E(A'B) + 2E(A'C) + 2E(B'C)$$

and, since $E(A'B) = E(A'C) = E(B'C) = 0$ and using that $E(\|B\|^2)$ and $E(\|C\|^2)$ are identical and iid, we obtain the final result. □

Having the previous proposition, we would like to calculate the last term of the error:

$$E(\|C\|^2) = E[(\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta^*)]$$

We can use the following *trick*:

$$E(\|C\|^2) = E[(\hat{\beta}-\beta)'X'X(\hat{\beta}-\beta^*)] = E(\|C\|^2) = E[tr((\hat{\beta}-\beta)'X'X(\hat{\beta}-\beta^*))] = E(\|C\|^2) = E[tr(XX'(\hat{\beta}-\beta^*)$$

and, recalling that $X(X'X)^{-1}X' = H$, we obtain that

$$tr(E\left[(X'X)(X'X)^{-1}X'\Sigma X(X'X)^{-1}\right]) = tr(\Sigma H).$$

We have just proved that

$$E_{\text{test}} = E_{\text{train}} + 2tr(\Sigma H).$$

Furthermore, recalling that

$$Cov(Y, \hat{Y}) = Cov(Y, HY) = \Sigma H$$

we obtain that

$$E_{\text{test}} = E_{\text{train}} + 2\sum_{i=1}^{n} Cov(Y_i, \hat{Y}_i).$$

An interpretation for this is that, if $Y_i \approx \hat{Y}_i$, then we are being pretty optimistic.

Case: Independence and homocedasticity

Consider that

$$\begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix} = \sigma^2 I$$

then,

$$tr(\Sigma H) = \sigma^2 tr(H) = \sigma^2 p$$

$$tr(X(X'X^{-1}X')) = p$$

Hence,

$$E_{\text{test}} = E_{\text{train}} + 2p\sigma^2$$

Case: Independence and heterocedastic XDXDXD

In this case,

$$\begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_n^2 \end{pmatrix} = \sigma^2 I$$

Hence,

$$E_{\text{test}} = E_{\text{train}} + 2\sigma_i^n \sigma_i^2 h_{ii}$$

Parte IV.
Classification