Advanced Statistical Modeling

Part 2. Nonparametric Modeling

Session 5:

Inference with nonparametric regression

Pedro Delicado

Departament d'Estadística i Investigació Operativa

Universitat Politècnica de Catalunya

Variability bands

Testing for no effects

Checking a parametric model

Comparing curves

Variability bands

- We have seen that the local linear estimator of the nonparametric regression model has bias
$$E(\hat{m}(x)) - m(x) = \frac{h^2 m''(x)\mu_2(K)}{2} + o(h^2)$$
and variance
$$V(\hat{m}(x)) = \frac{R(K)\sigma^2(x)}{nhf(x)} + o\left(\frac{1}{nh}\right).$$
- The variance can be estimated as
$$\hat{V}(\hat{m}(x)) = \frac{R(K)\hat{\sigma}^2(x)}{nhf(x)},$$
$\hat{\sigma}^2(x)$ being any estimate of conditional $V(Y|X=x) = \sigma^2(x)$.
- Assuming constant variance (homoscedastic model), then $\hat{\sigma}^2(x) = \hat{\sigma}^2$ for all x, where $\hat{\sigma}^2$ is one of the σ^2 estimators we have already studied.

- For a fix value of h it can be proved that

$$\frac{\hat{m}(x) - E(\hat{m}(x))}{\sqrt{V(\hat{m}(x))}} \longrightarrow N(0,1) \text{ in distribution as } n \longrightarrow \infty.$$

- This fact allows us to define (asymptotic) confidence intervals for $E(\hat{m}(x))$, that we are calling variability bands for $\hat{m}(x)$.
- Observe that they are not confidence intervals for $m(x)$.
- For $\alpha = 0.05$,

$$IC_{1-\alpha}(E(\hat{m}(x))) \equiv \left(\hat{m}(x) \mp 1.96 \sqrt{\hat{V}(\hat{m}(x))} \right).$$

- In generalized nonparametric regression models, each local model fitted by maximum local likelihood gives rise to an estimate of the local parameter $\hat{\theta}(x)$, and also an estimate of its variance $\hat{V}(\hat{\theta}(x))$.
- Taking into account that

$$\hat{m}(x) = \hat{E}(Y|X = x) = g^{-1}(\hat{\theta}(x))$$

  and using the delta method, it follows that an estimate of $V(\hat{m}(x))$ is

$$\hat{V}(\hat{m}(x)) = \hat{V}(\hat{\theta}(x)) / \left(g'(\hat{m}(x))\right)^2,$$

  where g' is the derivative of function g.
- Variability bands for $\hat{m}(x)$ can then be defined from $\hat{V}(\hat{m}(x))$ as in the standard nonparametric regression model.

- Observe that the (asymptotic) confidence $(1-\alpha)$ of variability bands is pointwise for $E(\hat{m}(x))$. They are not uniform confidence bands.
- A hypothetical uniform band for $E(\hat{m}(x))$ should be a pair of random functions $L(x)$ and $U(x)$ verifying that

$$P\left(L(x) \le E(\hat{m}(x)) \le U(x), \text{ for all } x \in \mathbb{R}\right) \approx 1-\alpha.$$

- See Section 5.7 of Wasserman (2006) for a way to compute uniform bands for $m(x)$.

Practice:

Variability bands

Variability bands

Testing for no effects

Checking a parametric model

Comparing curves

Testing for no effects

In the nonparametric regression model $Y_i = m(x_i) + \varepsilon_i$, $i = 1, \dots, n$, we test the null hypothesis of no effects:

$$\begin{cases} H_0 : m(x) \text{ is constant in } x \text{ and equal to } \mu_Y = E(Y), \\ H_1 : m(x) \text{ is not constant in } x. \end{cases}$$

Working analogously as in the multiple linear regression models, we use the test statistic

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS}_1)/(\mathrm{df}_0 - \mathrm{df}_1)}{\mathrm{RSS}_1/\mathrm{df}_1},$$

where the residual sums of squares (RSS_j) and the corresponding degrees of freedom (df_j) are $\mathrm{df}_0 = n - 1$,

$$\mathrm{RSS}_0 = \sum_{i=1}^{n} (y_i - \bar{y})^2, \; \mathrm{RSS}_1 = \sum_{i=1}^{n} (y_i - \hat{m}(x_i))^2,$$

$\hat{m}(x)$ is a nonparametric estimate with effective degrees of freedom df_1.

- The theoretical distribution of the test statistic F under the null hypothesis of no effects is unknown (it is known that it follows a F distribution in the linear model case with Gaussian residuals).
- The way the null distribution of F is tabulated in practice is by a permutation test.
- If H_0 is true, any permutation of $y_1, \ldots, y_n$ is equally likely for $x_1, \ldots, x_n$ fixed.
- Then the null distribution of F is approximated by the following algorithm.

Permutation test for the no effects test

1. Randomly permute $y_1, \ldots, y_n$ to obtain $y_{i_1}, \ldots, y_{i_n}$. Define the permuted sample as

$$(x_j, y_{i_j}),\ j = 1, \ldots, n.$$

2. Compute the value of the statistic F in the permuted sample: F_P.
3. Repeat B times steps 1 and 2: $F_P^1, \ldots, F_P^B$.
4. Compare the observed value of F in the original sample, F_{obs}, with $F_P^1, \ldots, F_P^B$, and obtain the test p-value:

$$p\text{-value} = \frac{\#\{F_P^b > F_{obs}\}}{B}.$$

Graphical reference band for the no effects model

- In Step 2 of the preceding permutation procedure, for each permuted sample a nonparametric estimation of the constant regression function has been done.
- Represent all these B estimated functions simultaneously at the same graphic to obtain a reference band for the no effects model.
- This reference band allows us to test graphically the null hypothesis of no effects:
  - If the estimated function is outside the reference band then reject the null hypothesis of no effects.

Alternative reference band for the no effects model

- We present here a different reference band for the no effects model that does not require the use of permuted samples.
- Under the null hypothesis of no effects ($m(x) = \mu_Y$, constant in x) the local linear estimator is unbiased:

$$\hat{m}(x) = \sum_{i=1}^{n} w^*(x_i, x) y_i \Rightarrow E(\hat{m}(x)) = \sum_{i=1}^{n} w^*(x_i, x)\mu_Y = \mu_Y = m(x).$$

- Let $\bar{y}$ be the sample mean of $y_1, \ldots, y_n$. This is also an unbiased estimator of μ_Y.
- Then, for all x, $E(\hat{m}(x) - \bar{y}) = 0$, and

$$V(\hat{m}(x)-\bar{y}) = V(\sum_{i=1}^{n} w^*(x_i, x)y_i - \sum_{i=1}^{n}(1/n)y_i) = \sigma^2 \sum_{i=1}^{n}(w^*(x_i, x)-(1/n))^2.$$

- Asymptotic normality implies that

$$\left(\bar{y} \mp 1.96 \sqrt{\hat{\sigma}^2 \sum_{i=1}^{n} (w^*(x_i, x) - (1/n))^2}\right)$$

  is an approximated reference band, with confidence 0.95, for the no effects model.

- A nonparametric estimation $\hat{m}(x)$ outside this band indicates that H_0 should be rejected.
- Take into account that a graphical test is useful mainly as a descriptive tool, and that it is much less accurate than a permutation test.

Effect of bandwidth choice on the test result

- In the previous testing procedure the bandwidth value h has been fixed in all the nonparametric estimations of $m(x)$ that have been done for different permuted samples.
- Therefore the test p-value and the test result can depend on the bandwidth h we use.
- It is recommended to draw a plot of pairs $(h, p\text{-value}(h))$.
- Such a plot shows whether the test result depends on h or not.
- This recommendation is valid for any hypothesis testing involving nonparametric curve estimations depending on a smoothing parameter.

Testing the linear regression model

- In the model $Y_i = m(x_i) + \varepsilon_i$, test the linear regression hypothesis:

$$\begin{cases} H_0 : m(x) = \beta_0 + \beta_1 x, \\ H_1 : m(x) \text{ is not linear.} \end{cases}$$

- Let Y be the vector of the n response data y_i, X be the matrix of explanatory variables (including the constant term), and $H = X(X^\mathsf{T} X)^{-1} X^\mathsf{T}$ be the hat matrix.
- Thus the fitted values and the residuals of the linear model are

$$\hat{Y}_L = HY,\ \hat{\varepsilon}_L = Y - \hat{Y}_L = (I_n - H)Y.$$

- Testing the linear model is equivalent to testing no effects in the relation between the estimated residuals $\hat{\varepsilon}_{L,i}$, and x_i:

$$\begin{cases} H_0 : E(\hat{\varepsilon}_{L,i}) = 0, \\ H_1 : E(\hat{\varepsilon}_{L,i}) = m(x_i) - (\beta_0 + \beta_1 x_i). \end{cases}$$

- In particular, an approximated reference band, with confidence 0.95, for the null hypothesis of linearity is

$$\left(\hat{\beta}_0 + \hat{\beta}_1 x \mp 1.96 \sqrt{\hat{\sigma}^2 \sum_{i=1}^{n} (w^*(x_i, x) - h(x_i, x))^2}\right),$$

where $h(x_i, x)$ is the i-th element of the row vector $h(x) = (1, x)(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}$, doing

$$\hat{y}_x = \hat{\beta}_0 + \hat{\beta}_1 x = (1, x)\hat{\beta} = (1, x)(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}Y = h(x)Y.$$

- If the nonparametric estimation $\hat{m}(x)$ is outside the reference band then H_0 should be rejected.

Practice:

Testing the linear model

Testing the Generalized Linear Model

- Consider the generalized nonparametric regression model

$$(Y|X = x) \sim f(y; m(x), \psi),$$

with link function g and $\theta(x) = g(m(x))$ a non-constrained smooth function of x.

- We want to test whether $\theta(x)$ is a linear function or not:

$$\begin{cases} H_0 : g(m(x)) = \beta_0 + \beta_1 x, \\ H_1 : g(m(x)) \text{ is not a linear function of } x. \end{cases}$$

- We use a pseudo-likelihood ratio test, an analogous test to likelihood ratio tests used when testing nested parametric models.
- The test statistic is

$$\mathrm{PLRT} = 2\sum_{i=1}^{n}\left(\log f(y;\hat{m}(x_i),\hat{\psi}_{NP}) - \log f(y;g^{-1}(\hat{\beta}_0+\hat{\beta}_1 x_i),\hat{\psi}_{GLM})\right),$$

where:

  - $\hat{m}(x)$ is a nonparametric estimate of $m(x)$ (possibly the maximum local likelihood estimator),
  - $\hat{\psi}_{NP}$ is the estimate of ψ derived from the nonparametric estimation of $m(x)$,
  - $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\psi}_{GLM}$ are the estimates provided by the GLM fitting.

- The null distribution of the test statistic PLRT is tabulated by parametric bootstrap, a procedure that allows us to generate samples according to the null hypothesis that are as similar as possible to the observed sample:
  1. Estimate the GLM from the observed data: $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\psi}_{GLM}$.
  2. Generate a bootstrap sample: for each value x_i, simulate y_i^* from the model
  $$(Y|X = x_i) \sim f(y; g(\hat{\beta}_0 + \hat{\beta}_1 x_i), \hat{\psi}_{GLM}).$$
  3. Compute the test statistic PLRT from the bootstrap sample: PLRT^*.
  4. Repeat B times steps 2 and 3: $\text{PLRT}_1^*, \dots, \text{PLRT}_B^*$.
  5. Compare the observed value of the test statistic PLRT at the original sample, PLRT_{obs}, with $\text{PLRT}_1^*, \dots, \text{PLRT}_B^*$, and obtain the test p-value:
  $$p\text{-value} = \frac{\#\{\text{PLRT}_b^* > \text{PLRT}_{obs}\}}{B}.$$
- It is possible to build a reference band for the GLM assumed under H_0 that provides a graphical test.

Practice:

- Testing the logistic model.
- Testing the Poisson GLM.

Variability bands

Testing for no effects

Checking a parametric model

Comparing curves

Testing equality of regression functions

- Let us assume that observed data come from I different subpopulations, and that they obey a possibly different nonparametric regression model at each one:

$$y_{ij} = m_i(x_{ij}) + \varepsilon_{ij},\ j = 1, \dots, n_i,\ i = 1, \dots, I.$$

- We want to test the equality of the I regression curves:

$$\begin{cases} H_0 : m_i(x) = m(x),\ i = 1, \dots, I,\ \text{for all } x, \\ H_1 : \text{not all the regression functions are equal.} \end{cases}$$

- A convenient test statistic is

$$T_I = \frac{\sum_{i=1}^{I} \sum_{j=1}^{n_i} (\hat{m}_i(x_{ij}) - \hat{m}(x_{ij}))^2}{\hat{\sigma}^2},$$

  - $\hat{m}(x)$ is the nonparametric estimate of $m(x)$ under the null hypothesis, that is, using all the observed data jointly;
  - $\hat{m}_i(x)$ is the nonparametric estimate of $m(x)$ using data from subpopulation i, $i = 1, \ldots, I$;
  - $\hat{\sigma}^2$ is the pooled estimated of $\sigma^2 = V(\varepsilon_{ij})$, defined from the estimates of σ^2 at each subpopulation,

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{I} \eta_i \hat{\sigma}_i^2}{\sum_{i=1}^{I} \eta_i},$$

    where η_i is the effective number of degrees of freedom in the estimation of $m(x)$ at the i-th subpopulation.
- Statistic T_I has the usual structure of ANOVA test statistics: between groups variation divided by within groups variation.

Tabulating the null distribution of T_I

- There are different alternative ways to tabulate the null distribution of T_I.
- Option 1: Permutation test.
  - If H_0 is true then the label indicating subpopulation can be interchanged between individuals without any alteration in the distribution of statistic T_I.
  - Thus B samples are generated by random permutation of subpopulation labels.
  - At each permuted sample the value of statistic T_I is computed and the values T_I^b, $b = 1, \ldots, B$, are obtained.
  - The test p-value is

$$p\text{-value} = \frac{\#\{T_I^b > T_{I,obs}\}}{B}.$$

Option 2: Bootstrap

1. Compute the residuals from the nonparametric estimation done at each subpopulation,
$$\hat{e}_{ij} = y_{ij} - \hat{m}_i(x_{ij}),\ j = 1, \ldots, n_i,\ i = 1, \ldots, I,$$
and define the set $E = \{\hat{e}_{ij},\ j = 1, \ldots, n_i,\ i = 1, \ldots, I\}$.
2. Generated a bootstrap sample as follows:
$$y^*_{ij} = \hat{m}(x_{ij}) + \hat{e}^*_{ij},\ j = 1, \ldots, n_i,\ i = 1, \ldots, I$$
where $\hat{e}^*_{ij}$ are randomly selected from set E with replacement.
3. Compute the statistic T_I at each bootstrap sample: T^*_I.
4. Repeat 2 and 3 B times steps: $T^*_{I,1}, \ldots, T^*_{I,B}$.
5. Define the test p-value:
$$p\text{-value} = \frac{\#\{T^*_{I,b} > T^*_{I,obs}\}}{B}.$$

Graphical test for two subpopulations

- For the two subpopulation case ($I = 2$) the preceding test can be complemented with an approximated graphical test.
- It consists of drawing a reference band around the global estimate $\hat{m}(x)$.
- If the null hypothesis is true, the estimation of $m(x)$ at both subpopulations should fall within the reference band.
- Under the null hypothesis, $d(x) = m_1(x) - m_2(x) = 0$ for all x. Let

$$\hat{d}(x) = \hat{m}_1(x) - \hat{m}_2(x)$$

  be the estimate of the difference function.
- Its variance is $V(\hat{d}(x)) = V(\hat{m}_1(x)) + V(\hat{m}_2(x))$ and it can be estimated following the ideas introduced when variability bands were developed.

- Finally, for $\alpha = 0.05$, the reference bands for the null hypothesis are

$$C(x) \equiv \left(\frac{1}{2}(\hat{m}_1(x) + \hat{m}_2(x)) \mp \frac{1,96}{2}\sqrt{\hat{V}(\hat{d}(x))} \right).$$

- It is easy to verify that

$$\hat{m}_1(x) \notin C(x) \iff \hat{m}_2(x) \notin C(x) \iff |\hat{d}(x)| > 1.96\sqrt{\hat{V}(\hat{d}(x))}.$$

- The reference bands are pointwise, they are not uniform bands.
- These reference bands suggest an alternative test statistic:

$$T_d = \int_{\mathbb{R}} \frac{(\hat{d}(x))^2}{\hat{V}(\hat{d}(x))} f(x)dx.$$

Its null distribution could be approximated using permuted or bootstrap samples.

Practice:

Comparing curves

Bowman, A. W. and A. Azzalini (1997).
Applied Smoothing Techniques for Data Analysis.
Oxford: Oxford University Press.

Hastie, T., R. Tibshirani, and J. Friedman (2001).
The Elements of Statistical Learning. Data Mining, Inference, and Prediction.
Springer.

Wasserman, L. (2006).
All of Nonparametric Statistics.
New York: Springer.