Lecture 25: Worked Examples

Sofia Olhede

EPFL

December 15, 2020

Worked examples

- I have been asked to cover:
- For instance over testing hypothesis on the last exercise sheet;
- Non–parametric regression again;
- More concrete examples on GLMs;
- How to handle the separable data case again, along with jitter residuals;
- Non–parametric regression;
- Estimate the unknown function $h(x)$ with the modulators and the wavelets.

Hypothesis testing in the LM (compare lec 15)

- First let us start with a model

$$\mathbb{E}\, Y_i = \beta_0 + \beta_1 x_i, \quad i = 1, \dots, n,$$

versus the more complex model

$$\mathbb{E}\, Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2, \quad i = 1, \dots, n.$$

To be able to have a full rank model $\mathrm{rank}\{\mathsf{X}\} = 3$ let us assume that $\overline{x}^2 \neq n^{-1} \sum x_i^2$. (What are we excluding?)

- How do we fit this model?
- We first write down the model

$$\mathsf{X} = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \dots & \dots & \dots \\ 1 & x_n & x_n^2 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}.$$

Hypothesis testing in the LM (compare lec 15)

- We then specify the A matrix as

$$A = \begin{pmatrix} 0 & 0 & 1 \end{pmatrix}.$$

- This will allow us to test if $\beta_3 = 0$ (is there a quadratic effect?)
- The simpler model then be fitted with

$$X_0 = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \dots \\ 1 & x_n \end{pmatrix}, \quad \boldsymbol{\beta}_0 = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}.$$

Testing in the Least Squares Set–Up

- Then

Source	d.o.f.	Sum of Squares	Mean squares	F
Red	$p-s$	$\underline{y}^T \mathrm{P}_0 \underline{y}$		
H_0	s	$\underline{y}^T(\mathrm{P}-\mathrm{P}_0)\underline{y}$	$M_1 = \frac{\underline{y}^T(\mathrm{P}-\mathrm{P}_0)\underline{y}}{s}$	$\frac{M_1}{M_2}$
Residual	$n-p$	$\underline{y}^T(\mathrm{I}-\mathrm{P})\underline{y}$	$M_2 = \frac{\underline{y}^T(\mathrm{I}-\mathrm{P})\underline{y}}{n-p}$	
total	n	$\underline{y}^T\underline{y}$		

- $M_2 = \frac{RSS}{n-p}$ is an *unbiased* estimate of σ^2.
- Reject the null hypothesis at level α if

$$F > f_\alpha$$

where

$$P(F_{s,n-p} > f_\alpha) = \alpha.$$

- This can be done even more easily using orthogonal polynomials, e.g. $z_{i1} = x_i - \bar{x}$ and $z_{i2} = (x_i - \bar{x})^2 + b(x_i - \bar{x}) + c$ where $b = -\sum_i \{x_i - \bar{x}\}^3 / \sum_j \{x_j - \bar{x}\}^2$ & $c = (-1/n)\sum_j \{x_j - \bar{x}\}^2$.

Another example

- Assume that

$$\mathbb{E}\{Y_i\} = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}, \quad i = 1, \ldots, 23.$$

- Here $p = 3$. Assume we are given

$$\mathrm{X}^T\mathrm{X} = \begin{pmatrix} 16 & 8 & 4 \\ 8 & 6 & 2 \\ 4 & 2 & 6 \end{pmatrix}, \quad \mathrm{X}^T\mathrm{y} = \begin{pmatrix} 100 \\ 60 \\ 40 \end{pmatrix}, \quad \mathrm{y}^T\mathrm{y} = 1240.$$

- Assume we wish to test $H_0 : \beta_2 = \beta_3 = 0$.
- Therefore the matrix A is given by

$$\mathrm{A} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Another example cont'd

- We can calculate $(X^TX)^{-1}$ and find it to be:

$$(X^TX)^{-1} = \begin{pmatrix} 0.2 & -0.25 & -0.05 \\ -0.25 & 0.5 & 0 \\ -0.05 & 0 & 0.2 \end{pmatrix}.$$

- With the additional assumption of NTA we can test. Combining the stated matrices we get

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} 3 \\ 5 \\ 3 \end{pmatrix}.$$

- We can calculate $\mathrm{Red}\{\widehat{\boldsymbol{\beta}}\} = \widehat{\boldsymbol{\beta}}^T X^T y = 720$.
- Furthermore $R^2 = y^Ty - \mathrm{Red}\{\widehat{\boldsymbol{\beta}}\} = 1240 - 720 = 520$.

Another example cont'd

- Under H_0 the model is

$$\mathbb{E}\{Y_i\} = \beta_1 x_{1i}, \mathrm{X}_0 = \begin{pmatrix} x_{1,1} \\ \dots \\ x_{1,23} \end{pmatrix}.$$

- Our estimated parameter is

$$\widehat{\boldsymbol{\beta}}_0 = \begin{pmatrix} 100/16 \\ 0 \\ 0 \end{pmatrix}.$$

- Then $\mathrm{Red}\{\widehat{\boldsymbol{\beta}}_0\} = 625$.
- So $R_0^2 = \mathrm{y}^T\mathrm{y} - \mathrm{Red}\{\widehat{\boldsymbol{\beta}}_0\} = 1240 - 625 = 615$.
- So $R_0^2 - R^2 = 615 - 520 = 95$.

Testing in the Least Squares Set–Up

- Then

Source	d.o.f.	Sum of Squares	Mean squares	F
Red	1	625		
H_0	2	95	47.5	1.83
Residual	20	520	26	
total	23	1240		

- Reject the null hypothesis at level 5 if

$$F > f_{1-\alpha} = 3.49$$

where we are looking at the $F_{2,20}$ distribution. $1.83 < 3.49$. So we cannot reject the null.

$$P(F_{s,n-p} \leq f_\alpha) = \alpha.$$

What is a LM?

- I thought we would also go back to basics a bit...
- What **is** a linear model?
- Is
$$\mathbb{E}\{Y_i\} = \beta_1 + \beta_2 x_i^{\beta_3}?$$
No.
- Is
$$\mathbb{E}\{Y_i\} = \beta_2 x_i^{\beta_3}?$$
Kind of. Take logarithms and assume we can handwave about the order of $\log(\cdot)$ and $\mathbb{E}$.
$$\mathbb{E}\{\log Y_i\} = \log \beta_2 + \beta_3 \log x_i.$$
- Do not want non–linear functions of the parameters $\boldsymbol{\beta}$.

What is a non–full rank LM?

- Let us try another one.
- Assume we are doing a twin study. We have 10 pairs of identical twins. We have

$$\mathbb{E}\{Y_{Ai}\} = \mu_i + \mu_A,\ i = 1, \dots, 10 \tag{1}$$
$$\mathbb{E}\{Y_{Bi}\} = \mu_i + \mu_B,\ i = 1, \dots, 10. \tag{2}$$

  In this example $p = 12$ as there are 12 parameters. We can write down

$$\mathrm{X} = \begin{pmatrix} 1 & 0 & 1 & 0 & \dots & 0 \\ 1 & 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 1 & 0 & 0 & \dots & 1 \end{pmatrix}.$$

- But the rank of X is $11 < 12$. We need to add a constraint to estimate the parameters. See linear dependence in the columns; add first two columns and subtract the sum of the last ones.

What is a non-full rank LM?

- We define the matrix H. In the previous example we might take

$$\mathsf{H} = \begin{pmatrix} 1 & 1 & 0 & 0 & \dots & 0 \end{pmatrix}.$$

- Thus we end up with with two equations:

$$\begin{cases} \mathsf{X}^T\mathsf{X}\widehat{\boldsymbol{\beta}} & = & \mathsf{X}^T\mathsf{y} \\ \mathsf{H}\widehat{\boldsymbol{\beta}} & = & 0 \end{cases}.$$

- If we multiply the second equation by H^T and add to the first then we get

$$\{\mathsf{X}^T\mathsf{X} + \mathsf{H}^T\mathsf{H}\}\widehat{\boldsymbol{\beta}} = \mathsf{X}^T\mathsf{y}.$$

- Our solution then becomes:

$$\widehat{\boldsymbol{\beta}} = \{\mathsf{X}^T\mathsf{X} + \mathsf{H}^T\mathsf{H}\}^{-1}\mathsf{X}^T\mathsf{y}.$$

What is a non-full rank LM?

- We can define special combination of β that can always be identified.
- These are called "estimable functions."

What about non-parametric regression?

- We saw before that a standard problem to solve was to estimate a function $g(x_i)$ submersed in noise.
- This is our model is

$$Y_i = h(x_i) + \epsilon_i, \quad i = 1, \ldots, n.$$

- We want to estimate $h(x_i)$

Orthogonal functions

- Suppose again that we observe

$$Y_i = h(x_i) + \varepsilon_i, \quad i = 1, \ldots, n.$$

- Here $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ are iid.
- Initially we assume $x_i = i/n$ namely a regular design for $i = 1, \ldots, n$.
- Let $\phi_1(x)$, $\phi_2(x), \ldots$ be an orthogonal basis for the interval $[0, 1]$. Often the cosine basis is used

$$\phi_1(x) = 1, \quad \phi_j(x) = \sqrt{2}\cos(\{j-1\}\pi x), \quad x = 2,\ 3\ldots.$$

- Here we expand $h(x)$ as

$$h(x) = \sum_{j=1}^{\infty} \theta_j \phi_j(x),$$

where $\theta_j = \int_0^1 h(x)\phi_j(x)\,dx$.

Orthogonal functions II

- We approximate

$$h_n(x) = \sum_{j=1}^{n} \theta_j \phi_j(x),$$

  which is a projection of $h(x)$ into the span of $\{\phi_1(x), \phi_2(x), \ldots, \phi_n\}$.
- This introduces an integrated squared bias of

$$B_n(\theta) = \int_0^1 \{r(x) - r_n(x)\}^2 \, dx = \sum_{j=n+1}^{\infty} \theta_j^2.$$

- We can understand this further.

Orthogonal functions III

- This can be quantified.
  Lemma: Let $\Theta(m, c)$ be a Sobolev ellipsoid. Then

$$\sup_{\theta \in \Theta(m,c)} B_n(\theta) = O\left(\frac{1}{n^{2m}}\right).$$

- A Sobolev ellipsoid is a set of functions for which $\theta_j^2 \sim (\pi j)^{2m}$; an ellipsoid is defined by

$$\Theta = \left\{ \theta : \sum_j a_j^2 \theta_j^2 \leq c^2 \right\}.$$

- Therefore if $m > 1/2$ we find $B_n = o(1/n)$.
- The bias is negligible and we shall ignore it for the rest of the chapter. We will therefore focus on estimating $h_n(x)$ rather than $h(x)$.

Orthogonal functions IV

- We define
$$Z_j = \frac{1}{n}\sum_{i=1}^{n} Y_i\phi_j(x_i), \quad j = 1, 2, 3, \ldots.$$
- We can then ask what is the distribution of Z_j?
- We note that
$$\begin{aligned} Z_j &= \frac{1}{n}\sum_{i=1}^{n} Y_i\phi_j(x_i) \\ &= \frac{1}{n}\sum_{i=1}^{n}\{h(x_i)+\varepsilon_i\}\phi_j(x_i) \\ &= \theta_j + \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\phi_j(x_i) = \theta_j + \nu_j. \end{aligned} \tag{3}$$

Using earlier results we can deduce that $\nu_j \sim N(0, \frac{\sigma^2}{n})$.

Orthogonal functions V

- We know from a previous section (Lecture 7) that shrinkage estimators can reduce the mean square error.
- We shall discuss James-Stein estimators a bit further.
- A modulator is a vector $b = \begin{pmatrix} b_1 & \dots & b_n \end{pmatrix}$ such that $0 \leq b_j \leq 1$ for $j = 1, \dots, n$.
- A modulation estimator takes the form

$$\begin{aligned} \widehat{\theta} &= b \odot Z \\ &= \begin{pmatrix} b_1 Z_1 \\ \dots \\ b_n Z_n \end{pmatrix}. \end{aligned} \tag{4}$$

- A constant modulator is a modulation of the form $\begin{pmatrix} b & \dots & b \end{pmatrix}$.
- A nested subset selection modulator is a modulator of the form $\begin{pmatrix} b & \dots & b & 0 & \dots & 0 \end{pmatrix}$.

Orthogonal functions VI

- A monotone modulator is of the form

$$1 \geq b_1 \geq b_2 \geq \cdots \geq b_n \geq 0.$$

- The function estimator provided by a modulator is

$$\hat{h}_n(x) = \sum_{j=1}^{n} \widehat{\theta}_j \phi_j(x) = \sum_{j=1}^{n} b_j Z_j \phi_j(x).$$

  This is a linear smoother.
- Modulators shrink Z_j towards 0. This smoothes the function estimates.
- We define the risk as

$$R(b) = \mathbb{E}_\theta \{ \sum_{j=1}^{n} (b_j Z_j - \theta_j)^2 \}$$

Orthogonal functions VII

- To estimate b we need to estimate σ. There are reasons why we would take
$$\widehat{\sigma}^2 = \frac{1}{n - J_n} \sum_{i=n-J_n+1}^{n} Z_i^2.$$
- Often we take $J_n = n/4$.
- Theorem: The risk of a modulator b is
$$R(b) = \sum_{j=1}^{n} \theta_j^2 (1 - b_j)^2 + \frac{\sigma^2}{n} \sum_{j=1}^{n} b_j^2.$$
- The SURE estimator of $R(b)$ are
$$\widehat{R}(b) = \sum_{j=1}^{n} \left(Z_j^2 - \frac{\widehat{\sigma}^2}{n} \right)_+ (1 - b_j)^2 + \frac{\widehat{\sigma}^2}{n} \sum_{j=1}^{n} b_j^2.$$

Orthogonal functions VIII

- The modulation estimator of θ is

$$\theta = \left(\widehat{b}_1 Z_1,\ \widehat{b}_2 Z_2,\ \ldots\right).$$

where b minimises $\widehat{R}(b)$. This yields

$$\hat{h}_n(x) = \sum_{j=1}^{n} \widehat{\theta}_j \phi_j(x) = \sum_{j=1}^{n} b_j Z_j \phi_j(x).$$

For a fixed b we expect that $\widehat{R}(b)$ approximates $R(b)$. We need more, as $\hat{b}$ will depends on the same data as $\hat{R}(b)$. We therefore need $\widehat{R}(b)$ to approximate $R(b)$ uniformly.

- We shall assume that the modulator takes the form

$$\begin{pmatrix}1 & \ldots 1 & 0 & \ldots & 0\end{pmatrix}.$$

Orthogonal functions IX

- This corresponds to picking J to minimize

$$\widehat{R}(J) = \frac{J\widehat{\sigma}^2}{n} + \sum_{j=J+1}^{n} \left(Z_j^2 - \frac{\widehat{\sigma}^2}{n} \right)_+ .$$

- We note that $\widehat{R}(b)$ is

$$\widehat{R}(b) = \sum_{i=1}^{n} \{b_i - g_i\}^2 Z_i^2 + \frac{\widehat{\sigma}^2}{n} \sum_{i=1}^{n} g_i .$$

- Here

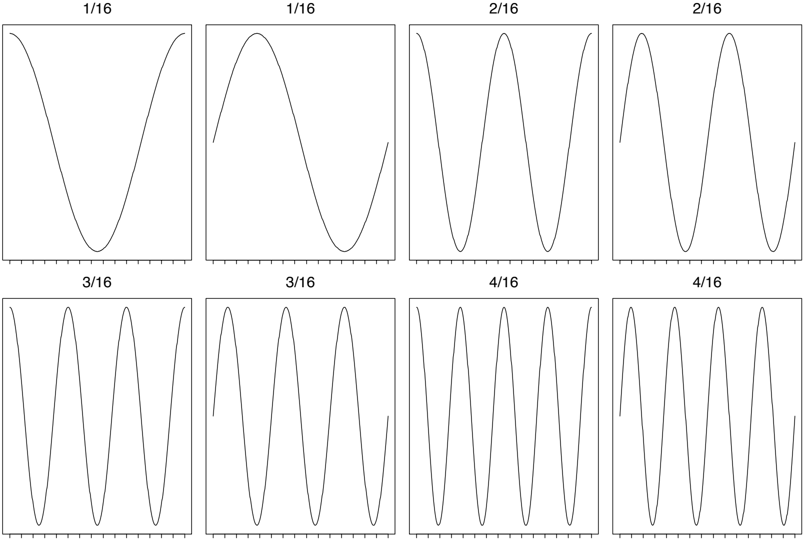
$$g_i = \{Z_i^2 - \frac{\widehat{\sigma}^2}{n}\}/Z_i^2 .$$

  We therefore minimize $\sum_{i=1}^{n}\{b_i - g_i\}^2 Z_i^2$.

Orthogonal functions X

- This then produces an estimator.
- The first generalization of this problem uses a basis that is orthogonal with respect to the design points $x_1, \ldots, x_n$.
- We define
$$Z_j = \frac{1}{n} \sum_{i=1}^{n} Y_i \phi(x_i).$$
- We can still use the developed methodology.
- The GLM version simply is not based on least squares.

Cosines & Sines



Cosines & Sines II