Data-based Statistical Decision Model

Lecture 8 Supplement - Inference for linear smoothers

Sungkyu Jung

Recall the linear smoothers

- Given samples $(x_i, y_i), i = 1, \ldots, n$, recall that a linear smoother is an estimator for the underlying regression function $m(x)$ satisfying

$$m(x_0) = \sum_{i=1}^{n} w(x_0, x_i) \cdot y_i = \mathbf{w}(x_0)'\mathbf{y},$$

at an arbitrary point x_0.

- Linear smoothers include the k-nearest neighbor regression, kernel regression, LOESS, smoothing splines and, of course, linear regression.

Linear Regression

- Fitted value at any x_0:

$$\hat{m}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0 = x_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\mathbf{y}.$$

- Fitted values at all observation points $x_1, \ldots, x_n$

$$\hat{\mathbf{y}} = \begin{bmatrix} \hat{m}(x_1) \\ \vdots \\ \hat{m}(x_n) \end{bmatrix} = \begin{bmatrix} x_1'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\mathbf{y} \\ \vdots \\ x_n'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\mathbf{y} \end{bmatrix} = \mathbf{X}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\mathbf{y} = \mathbf{H}\mathbf{y}.$$

Cubic Spline Smoothing

- Fitted value at any x_0: For $g(x_0) = (g_1(x_0), \ldots g_n(x_0))'$,

$$\hat{m}(x_0) = g(x)'\hat{\beta}_\lambda = g(x_0)'(\mathbf{G}'\mathbf{G} + \lambda\Omega)^{-1}\mathbf{G}'\mathbf{y}.$$

- Fitted values at all observation points $x_1, \ldots, x_n$: Since

$$\mathbf{G} = \begin{bmatrix} g(x_1) \\ \vdots \\ g(x_n) \end{bmatrix},$$

$$\hat{\mathbf{y}} = \begin{bmatrix} \hat{m}(x_1) \\ \vdots \\ \hat{m}(x_n) \end{bmatrix} = \mathbf{G}(\mathbf{G}'\mathbf{G} + \lambda\Omega)^{-1}\mathbf{G}'\mathbf{y}. = \mathbf{S}\mathbf{y}.$$

Review of inference in linear regression

- Assume $\mathbf{y} = \mathbf{X}\beta + \epsilon$ is correct, and $\epsilon \sim N(0, \sigma^2 \mathbf{I})$.
- We are implicitly assuming that $\beta = (\beta_0, \beta_1, \ldots, \beta_{p-1})$.

Pointwise confidence intervals

- $\hat{m}(x_0)$ has mean

$$E(\hat{m}(x_0)) = x_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}E(\mathbf{y}) = x_0'\beta = m(x_0).$$

- $\hat{m}(x_0)$ has variance

$$Var(\hat{m}(x_0)) = \sigma^2 x_0'(\mathbf{X}'\mathbf{X})^{-1}x_0.$$

  - Variance of $\hat{m}(x_i)$ at all observation points:

$$Var(\hat{\mathbf{y}}) = Var(\mathbf{H}\mathbf{y}) = \sigma^2 \mathbf{H}\mathbf{H}' = \sigma^2 \mathbf{H}.$$

- Estimate σ^2 by

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-p},$$

  where p here is the degrees of freedom of the model. Let

$$s^2(\hat{m}(x_0) = \hat{\sigma}^2 x_0'(\mathbf{X}'\mathbf{X})^{-1}x_0.$$

- Assuming normality of ϵ, we have

$$\frac{\hat{m}(x_0) - m(x_0)}{s(\hat{m}(x_0))} \sim t_{n-p}.$$

- Confidence interval for $E(Y \mid X = x_0) = m(x_0)$:

$$\hat{m}(x_0) \pm t_{n-p,\alpha/2} s(\hat{m}(x_0)),$$

where $t_{n-p,\alpha/2}$ is the $1-(\alpha/2)$ quantile of the t_{n-p} distribution.

- Prediction interval for $Y \mid X = x_0$ is

$$\hat{m}(x_0) \pm t_{n-p,\alpha/2} s_p(x_0),$$

where $s_p^2(x_0) = s^2(\hat{m}(x_0) + \hat{\sigma}^2$.

F-test for two nested linear models

Let M_1 and M_2 be two nested models, where M_1 has p_1 covariates, and M_2 has p_2 covariates, including all of p_1 covaraites in M_1.

```
data(Boston)
M1 <- lm(medv ~ lstat ) # M_1 with p_1 = 2 (including the intercept)
M2 <- lm(medv ~ lstat + I(lstat^2) + crim) # M_2 with p_2 = 4 (including the intercept)
```

Recall F-test for null model vs simple linear model

	M_1 **with** $p_1 = 1$	M_2 **with** $p_2 = 2$
	Null model	Linear model
Model	$Y = \beta_0 + \epsilon$	$Y = \beta_0 + \beta_1 X + \epsilon.$
Residuals	$y_i - \bar{y}$	$e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$
Sum of squares	$\sum_{i=1}^n (y_i - \bar{y})^2$	$\sum_{i=1}^n e_i^2$

Let

$$RSS_1 = \sum_{i=1}^n (y_i - \bar{y})^2, \quad RSS_2 = \sum_{i=1}^n e_i^2$$

or

$$RSS_1 = \sum_{i=1}^n (y_i - \hat{y}_i^{(1)})^2, \quad RSS_2 = \sum_{i=1}^n (y_i - \hat{y}_i^{(2)})^2.$$

Recall that the F-statistic is

$$\frac{(RSS_1 - RSS_2)/(p_2 - p_1)}{RSS_2/(n - p_2)}.$$

Source	df	SS	MS	F	p-value
Regression	1	$\mathrm{SS_{reg}}$	$\mathrm{MS_{reg}} = \frac{\mathrm{SS_{reg}}}{1}$	$F = \frac{\mathrm{MS_{reg}}}{\mathrm{MS_{res}}}$	
Residual	n-2	RSS	$\widehat{\sigma}^2 = \frac{\mathrm{RSS}}{n-2}$		
Total	n-1	$\mathrm{SS_{total}}$			

Test

Test the null hypothesis

$$H_0 : \beta_i = 0 \text{ for all } \ i \in M_2 \setminus M_1$$

versus

$$H_1 : \beta_i \neq 0 \text{ for some } \ i \in M_2 \setminus M_1.$$

If the null was true, then the F-statistic should follow

$$F = \frac{(RSS_1 - RSS_2)/(p_2 - p_1)}{RSS_2/(n - p_2)} \sim F_{p_2 - p_1, n - p_2}.$$

If the null was not true, then the value of F would be too large.

Inference with linear smoothers

- Just as in the linear regression case, we have

$$\hat{m}(x_0) = \sum_{i=1}^{n} w(x_0, x_i) \cdot y_i = \mathbf{w}(x_0)'\mathbf{y}$$

in the prediction of $m(x_0) = \mathrm{E}(Y \mid X = x_0)$. For smoothing splines, we may write $\hat{\mathbf{y}} = \mathbf{S}\mathbf{y}$.

- Let us assume $Y = m(X) + \epsilon$ is correct.
- Since we do not say anything about $m(\cdot)$, our smoother may not estimate m very well.

Variance of linear smoothers

- Just as in the linear regression,

$$Var(\hat{m}(x_0)) = Var(\mathbf{w}(x_0)'\mathbf{y}) = \sigma^2 \mathbf{w}(x_0)'\mathbf{w}(x_0).$$

  - Variance of $\hat{m}(x_i)$ at all observation points:

$$Var(\hat{\mathbf{y}}) = Var(\mathbf{S}\mathbf{y}) = \sigma^2 \mathbf{S}\mathbf{S}'.$$

- How to estimate σ^2? Use

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-d},$$

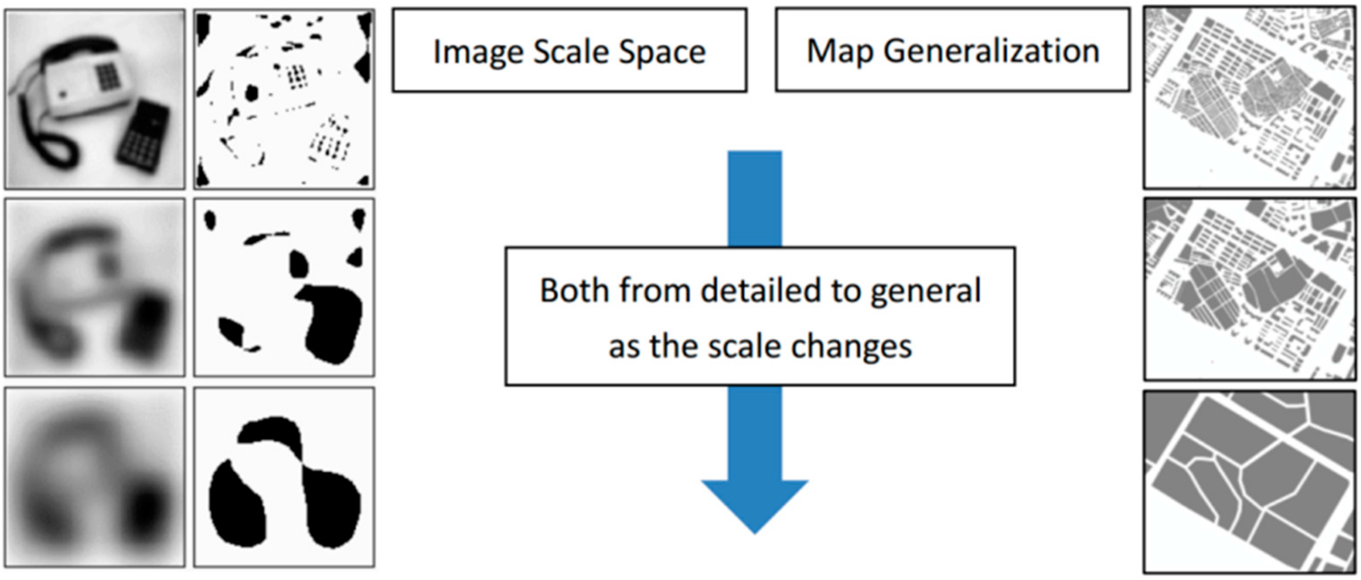
where d here is the degrees of freedom of the model.

  - d is not the number of parameters. (Recall in Smoothing Splines, the number of parameters is n!)
  - d describes the effective number of parameters used by linear smoothers.
  - For now, we set $d = \mathrm{trace}(\mathbf{S}) = \sum_{i=1}^{n} \mathbf{S}_{ii}$.
- Let

$$s^2(\hat{m}(x_0)) = \hat{\sigma}^2 \mathbf{w}(x_0)'\mathbf{x}(x_0).$$

Mean of linear smoothers: There is a bias!

- $E(\hat{m}(x_0)) = \mathbf{w}(x_0)'E(\mathbf{y}) = \sum_{i=1}^{n} w(x_0, x_i)m(y_i)$
- Predictor is indeed biased: $E(\hat{m}(x_0)) \neq m(x_0)$.
- Understand $\sum_{i=1}^{n} w(x_0, x_i)m(y_i)$ as a smoothed version of “true” m.



https://doi.org/10.3390/ijgi7020041 (https://doi.org/10.3390/ijgi7020041)

Implication on inference

- Assuming that $Y = m(X) + \epsilon$ is correct, and $\epsilon \sim N(0, \sigma^2)$, we have

$$\frac{\widehat{m(x_0)} - E(\widehat{m(x_0)})}{s(\hat{m}(x_0))} \sim t_{n-d},$$

  approximately.

- Our confidence interval

$$\hat{m}(x_0) \pm t_{n-d,\alpha/2}\, s(\hat{m}(x_0)),$$

captures the “smoothed true mean” $E(\widehat{m(x_0)})$ (not the unsmoothed, potentially jagged true mean $m(x_0) = E(Y \mid X = x)$) with proportion $1 - \alpha$.

Significance tests between fitted models

We will discuss an analogue of the F test in linear regression.

Suppose that we have two estimates $\hat{m}_1$ and $\hat{m}_2$ and the model class for $\hat{r}_1$ is nested within that of $\hat{m}_2$.

- Write $\hat{\mathbf{y}}^{(1)} = S_1\mathbf{y}$, $\hat{\mathbf{y}}^{(2)} = S_2\mathbf{y}$,

$$RSS_1 = \sum_{i=1}^{n}(y_i - \hat{y}_i^{(1)})^2, \quad RSS_2 = \sum_{i=1}^{n}(y_i - \hat{y}_i^{(2)})^2,$$

$$d_1 = \text{trace}(S_1), \quad d_2 = \text{trace}(S_2).$$

An example

A standard example is where $\hat{m}_1$ is a linear fit, $\hat{m}_2$ is a more flexible fit coming from, say, smoothing splines.

```
mod1 = lm(medv ~ lstat, data = Boston)
mod2 = smooth.spline(x = Boston$lstat,y = Boston$medv, df=100)
```

- In this case, $\mathbf{S}_1 = \mathbf{H}$ and $d_1 = \mathrm{trace}(\mathbf{H}) = 2$.
- Expressing the true regression function as

  $$Y = \beta_0 + \beta_1 X + \delta(X) + \epsilon,$$

  we can test the null hypothesis

  $$H_0 : \delta(x) = 0$$

  versus

  $$H_1 : \delta(x) \neq 0.$$

If the null was true, then the F-statistic should follow

$$F = \frac{(RSS_1 - RSS_2)/(d_2 - d_1)}{RSS_2/(n - d_2)} \sim F_{d_2 - d_1, n - d_2}.$$

If the null was not true, then the value of F would be too large.