Direct Inference with Linear Smoothers

Advanced Methods for Data Analysis (36-402/36-608)

Spring 2014

1 Review of 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 satisfying

$$\hat{r}(x_0) = \sum_{j=1}^{n} w(x_0, x_j) \cdot y_j,$$

at an arbitrary point x_0. We can alternatively express this as

$$\hat{r}(x_0) = w(x_0)^T y,$$

where $w(x_0) = (w(x_0, x_1), w(x_0, x_2), \ldots w(x_0, x_n)) \in \mathbb{R}^n$

- This means that the fitted value at the point x_i, $\hat{y}_i = \hat{r}(x_i)$, can be expressed as

$$\hat{y}_i = w(x_i)^T y,$$

and we can write the vector of fitted values $\hat{y} = (\hat{y}_1, \ldots \hat{y}_n)$ as

$$\hat{y} = Sy,$$

for the matrix $S \in \mathbb{R}^{n \times n}$ with ith row $w(x_i)$ (i.e., with $S_{ij} = w(x_i, x_j)$)

- We've seen that, e.g., linear regression, k-nearest neighbors regression, kernel regression, and smoothing splines are all linear smoothers. Important note: the latter three estimators are linear smoothers at any *fixed* value of their tuning parameters (k, h, and λ, respectively). Hence, e.g., if you want to think about smoothing splines as your linear smoother of choice, then just consider the smoothing parameter λ to be fixed at some value

2 Review of inference in linear regression

- For a particular type of linear smoother, namely, linear regression, we have a well-developed theory for statistical inference

- Suppose now that $x_i \in \mathbb{R}^p$, $i = 1, \ldots n$, and let x be the predictor matrix of dimension $n \times p$ (i.e., with ith row x_i). Define the usual regression coefficients $\hat{\beta} = (x^T x)^{-1} x^T y$. Then the fitted value at an arbitrary point $x_0 \in \mathbb{R}^p$ is

$$\hat{y}_{x_0} = x_0^T \hat{\beta} = x_0^T (x^T x)^{-1} x^T y.$$

Note that here $w(x_0) = x(x^T x)^{-1} x_0$, and

$$\hat{y} = Hy = x(x^T x)^{-1} x^T y$$

- We'll review inference for the fitted values $\hat{y}_{x_0} = x_0^T \hat{\beta}$. (As you learned, inference for the coefficients $\hat{\beta}$ is also possible, but we'll focus on the fitted values because this is what is relevant for the general setting of regression function prediction.) Assume that we observe

$$y_i = \beta^T x_i + \epsilon_i, \quad \epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2), \quad i = 1, \ldots n,$$

where x_i, $i = 1, \ldots n$ are considered fixed

2.1 Pointwise confidence intervals for the regression function

- Note that, at an arbitrary point x_0,

$$\begin{aligned} \mathrm{Var}(\hat{y}_{x_0}) &= \mathrm{Var}\big(w(x_0)^T y\big) \\ &= w(x_0)^T \mathrm{Var}(y) w(x_0) \\ &= \sigma^2 w(x_0)^T w(x_0) \\ &= \sigma^2 x_0^T (x^T x)^{-1} x_0 \end{aligned}$$

- Typically we must estimate σ^2 because it is unknown, and we use the residual sum of squares from the regression fit, as in

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

We know that $(n-p)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-p}$

- This yields the estimated variance of $\hat{y}_{x_0}$

$$\hat{s}^2(\hat{y}_{x_0}) = \hat{\sigma}^2 x_0^T (x^T x)^{-1} x_0,$$

and therefore

$$\frac{\hat{y}_{x_0} - \beta^T x_0}{\hat{s}(\hat{y}_{x_0})} \sim t_{n-p},$$

where t_{n-p} denotes a t distribution with $n-p$ degrees of freedom. To get a $(1-\alpha)$ confidence interval for the true value of the regression function $r(x_0) = \beta^T x_0$, hence, we use that

$$\begin{aligned} 1 - \alpha &= \mathbb{P}\Big(q_1 \leq \frac{\hat{y}_{x_0} - \beta^T x_0}{\hat{s}(\hat{y}_{x_0})} \leq q_2\Big) \\ &= \mathbb{P}\Big(\hat{y}_{x_0} - q_2 \hat{s}(\hat{y}_{x_0}) \leq \beta^T x_0 \leq \hat{y}_{x_0} - q_1 \hat{s}(\hat{y}_{x_0})\Big), \end{aligned}$$

where q_1, q_2 are the $\alpha/2$ and $(1-\alpha/2)$ quantiles of t_{n-p}, respectively

- I.e., $[\hat{y}_{x_0} - q_2 \hat{s}(\hat{y}_{x_0}), \hat{y}_{x_0} - q_1 \hat{s}(\hat{y}_{x_0})]$ is a $(1-\alpha)$ confidence interval for $r(x_0) = \beta^T x_0$. This is referred to as a *pointwise* confidence interval (emphasizing the fact that it guarantees coverage for the regression function at a single point x_0)

- Often, we'll want to construct a confidence interval for the underlying regression function over the observed input points, $r(x_i) = \beta^T x_i$, $i = 1, \ldots n$. From the above, we know that the ith such confidence interval is given by $[\hat{y}_i - q_2 \hat{s}(\hat{y}_i), \hat{y}_i - q_1 \hat{s}(\hat{y}_i)]$. Note that the (estimated) variance of $\hat{y}_i$ is here

$$\hat{s}^2(\hat{y}_i) = \hat{\sigma}^2 x_i^T (x^T x)^{-1} x_i.$$

Another way of looking at things, in matrix notation:

$$\mathrm{Var}(\hat{y}) = \mathrm{Var}(Hy) = \sigma^2 H H^T = \sigma^2 H,$$

and therefore $\mathrm{Var}(\hat{y}_i) = \sigma^2 H_{ii}$, and the estimated variance is $\hat{s}^2(\hat{y}_i) = \hat{\sigma}^2 H_{ii}$

2.2 Significance tests between fitted models

- We can also test for significance between two fitted nested regression models. Let $M_1 \subseteq M_2 \subseteq \{1, \ldots p\}$ be two nested sets, with sizes $p_1 = |M_1|$ and $p_2 = |M_2|$. Let $\hat{y}^{(1)}$ denote the vector of fitted values from the regression on variables in M_1, and $\hat{y}^{(2)}$ from the regression on variables in M_2. Define

$$\mathrm{RSS}_1 = \sum_{i=1}^{n} (y_i - \hat{y}_i^{(1)})^2, \quad \mathrm{RSS}_2 = \sum_{i=1}^{n} (y_i - \hat{y}_i^{(2)})^2,$$

the residual sum of squares from these two regressions. Then

$$\frac{(\mathrm{RSS}_1 - \mathrm{RSS}_2)/(p_2 - p_1)}{\mathrm{RSS}_2/(n - p_2)}$$

is the F statistic for testing the significance of variables in $M_2 \setminus M_1$, i.e., for testing the null hypothesis

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

versus the alternative

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

Under the null hypothesis, we have

$$\frac{(\mathrm{RSS}_1 - \mathrm{RSS}_2)/(p_2 - p_1)}{\mathrm{RSS}_2/(n - p_2)} \sim F_{p_2 - p_1, n - p_2},$$

where $F_{p_2 - p_1, n - p_2}$ denotes an F distribution with $(p_2 - p_1, n - p_2)$ degrees of freedom. Hence the test rejects for values of the statistic that exceed q, the $(1 - \alpha)$ quantile of $F_{p_2 - p_1, n - p_2}$

3 Inference with linear smoothers

3.1 Setup, the bootstrap, fixed versus random inputs

- Now we will learn the analogs of the above tools—pointwise confidence intervals, and F tests between fitted models—for general linear smoothers, beyond linear regression. Like the linear regression case, we will assume a model

$$y_i = r(x_i) + \epsilon_i, \quad \epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2), \quad i = 1, \ldots n,$$

where x_i, $i = 1, \ldots n$ are considered fixed. Because our estimator $\hat{r}$ is a linear smoother, we can write the fit as $\hat{r}(x_0) = w(x_0)^T y$ at an arbitrary point x_0, and $\hat{y} = Sy$ for the vector of fitted values across $x_1, \ldots x_n$

- To preface, there are certainly other ways to construct confidence intervals and significance tests than the "direct" ones we describe below. For example, we have already learned how to use the bootstrap in detail, and the bootstrap could be applied for both of these purposes. But the direct tools are more computationally efficient, have a close tie to those from linear regression, and are already implemented in R software, so they're worth knowing

- When mixing and matching tools, one thing to be aware of is the underlying assumption on the inputs $x_1, \ldots x_n$. The tools that we will describe below, just like those for linear regression, assume that these inputs are fixed. The standard pairs bootstrap, on the other hand, treats the inputs as random (since we resample pairs (x_i, y_i)). To use the boostrap and respect the fixed input setup, we'd have to use the residual bootstrap

- This is not to say that one route is generally less correct than the other, but rather, that these differences should be kept in mind when comparing the results produced by different tools. E.g., if we are comparing a pointwise confidence interval from linear regression, constructed via standard the methodology in the last section, to another one from a different estimator, constructed using the pairs bootstrap, we should be aware of the fact that these two tools are not actually considering the same level of randomness. (We would expect confidence intervals that are constructed in a random input setting to be generally wider, because they also incorporate the variability in $x_1, \ldots x_n$)

3.2 Pointwise confidence intervals for the regression function

- Just as in the linear regression case, at an arbitrary point x_0, the variance of the fit $\hat{r}(x_0) = w(x_0)^T y$ is

$$\mathrm{Var}\big(\hat{r}(x_0)\big) = \sigma^2 w(x_0)^T w(x_0)$$

- How to estimate σ^2? We can now use the estimate

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

where $d = \mathrm{df}(\hat{y}) = \mathrm{tr}(S)$, the degrees of freedom of the fit $\hat{y}$. Note: this replaces p in the usual expression for the estimated error variance in linear regression, so it should make intuitive sense to you from what you know about degrees of freedom. Now, $(n-d)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-d}$, but this is only an approximation in the case of general linear smoothers, and not exact like it was for linear regression. It is a good approximation nonetheless

- This yields the estimated variance of $\hat{r}(x_0)$

$$\hat{s}^2\big(\hat{r}(x_0)\big) = \hat{\sigma}^2 w(x_0)^T w(x_0),$$

and from the same arguments as before, an approximate $(1-\alpha)$ confidence interval for $r(x_0)$, the underlying regression function at a point x_0, is $[\hat{r}(x_0) - q_2 \hat{s}(\hat{r}(x_0)), \hat{r}(x_0) - q_1 \hat{s}(\hat{r}(x_0))]$, where q_1, q_2 are the $\alpha/2, (1-\alpha/2)$ quantiles of t_{n-d}, respectively

- For confidence intervals of the regression function at the observed inputs x_i, $i = 1, \ldots n$, the same story holds; an approximate confidence interval for $r(x_i)$ is $[\hat{y}_i - q_2 \hat{s}(\hat{y}_i)), \hat{y}_i - q_1 \hat{s}(\hat{y}_i))]$. Now

$$\hat{s}^2(\hat{y}_i) = \hat{\sigma}^2 w(x_i)^T w(x_i),$$

or another way of writing this is to use the fact that

$$\mathrm{Var}(\hat{y}) = \mathrm{Var}(Sy) = \sigma^2 S S^T,$$

so $\mathrm{Var}(\hat{y}_i) = \sigma^2 (SS^T)_{ii}$, and the estimated variance is $\hat{s}^2(\hat{y}_i) = \hat{\sigma}^2 (SS^T)_{ii}$

3.3 Learning to love the bias

- It is important to take a step back and think about the bias. Note that the confidence intervals in the last section utilize the t_{n-d} distribution for the calculation of quantiles q_1, q_2. As in the linear regression case, this stems from claiming that

$$\frac{\hat{r}(x_0) - r(x_0)}{\hat{s}(\hat{r}(x_0))} \sim t_{n-d}. \tag{1}$$

Remember that this is now an approximate result, because the denominator is only approximately χ^2_{n-d} (times constant factors). But there is something else going on too: in modeling this statistic as t_{n-d}, we are assuming that its numerator has mean zero, i.e.,

$$\mathbb{E}[\hat{r}(x_0)] = r(x_0),$$

or at least approximately so. Note that this is the same as saying that $\hat{r}(x_0)$ has zero bias or at least small bias. When this is true, i.e., when the bias is small, then we are more or less justified in saying that (2) holds, so that our confidence interval provides appropriate coverage for $r(x_0)$

- But when this is not true, i.e., when $\hat{r}(x_0)$ is badly biased, then we nevertheless have that

$$\frac{\hat{r}(x_0) - \mathbb{E}[\hat{r}(x_0)]}{\hat{s}(\hat{r}(x_0))} \sim t_{n-d}, \tag{2}$$

(or again, at least approximately so) and therefore the confidence interval that we construct $[\hat{r}(x_0) - q_2\hat{s}(\hat{r}(x_0)), \hat{r}(x_0) - q_1\hat{s}(\hat{r}(x_0))]$ is actually a confidence interval for $\mathbb{E}[\hat{r}(x_0)]$, rather than $r(x_0)$. So what is $\mathbb{E}[\hat{r}(x_0)]$? Well

$$\mathbb{E}[\hat{r}(x_0)] = w(x_0)^T r(x_0),$$

which is a smoothed version of $r(x_0)$. In terms of the fitted values $\hat{y} = (\hat{y}_1, \ldots \hat{y}_n)$, we have

$$\mathbb{E}[\hat{y}] = Sr,$$

where r is the vector of true regression function evaluations $r = (r(x_1), \ldots r(x_n))$, and again we can think of $\mathbb{E}[\hat{y}]$ as a smoothed version of the true regression function values

- Hence, in the presence of nonnegligible bias, we have to keep it in mind that our confidence intervals are really for $\mathbb{E}[\hat{r}(x_0)]$ or $\mathbb{E}[\hat{y}_i]$, which are smoothed versions of the true regression function values $r(x_0)$ and $r(x_i)$, and not the true values themselves

3.4 Significance tests between fitted models

- Here we present an analog of the F test in linear regression. Suppose that we are comparing two estimates $\hat{r}_1$ and $\hat{r}_2$, and the model class for $\hat{r}_1$ is nested within that of $\hat{r}_2$. Write

$$\hat{y}^{(1)} = S_1 y, \quad \hat{y}^{(2)} = S_2 y,$$

for the fitted values from $\hat{r}_1$ and $\hat{r}_2$ respectively,

$$d_1 = \mathrm{tr}(S_1), \quad d_2 = \mathrm{tr}(S_2),$$

for their respective degrees of freedom, and also

$$\mathrm{RSS}_1 = \sum_{i=1}^n (y_i - \hat{y}_i^{(1)})^2, \quad \mathrm{RSS}_2 = \sum_{i=1}^n (y_i - \hat{y}_i^{(2)})^2,$$

for their respective residual sums of squares

- A standard example is when $\hat{r}_1$ is a linear fit and $\hat{r}_2$ is a more flexible fit coming from (say) a smoothing spline. Expressing the true regression function as $r(x) = \beta_0 + \beta_1 x + \delta(x)$, we wish to test the null hypothesis

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

versus the alternative hypothesis

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

- In general, we must assume that $\hat{y}_i^{(2)} = \hat{r}_2(x_i)$ is approximately unbiased for $r(x_i)$, $i = 1, \ldots n$, and that $\hat{y}_i^{(1)} = \hat{r}_1(x_i)$ is approximately unbiased for $r(x_i)$, $i = 1, \ldots n$ under the null hypothesis. Then the F statistic for testing the significance of the fit $\hat{y}^{(2)}$ over $\hat{y}^{(1)}$ is

$$\frac{(\mathrm{RSS}_1 - \mathrm{RSS}_2)/(d_2 - d_1)}{\mathrm{RSS}_2/(n - d_2)},$$

and the null hypothesis, it holds that, approximately,

$$\frac{(\mathrm{RSS}_1 - \mathrm{RSS}_2)/(d_2 - d_1)}{\mathrm{RSS}_2/(n - d_2)} \sim F_{d_2 - d_1, n - d_2}.$$

As before, we reject when this statistic exceeds q, the $(1 - \alpha)$ quantile of $F_{d_2 - d_1, n - d_2}$

3.5 Additive models

- With additive models fit by backfitting, using linear smoothers, everything follows similarly. Suppose now that each $x_i = (x_{i1}, \ldots x_{ip}) \in \mathbb{R}^p$, $i = 1, \ldots n$ and we fit the additive model

$$\hat{r}(x_i) = \hat{r}_1(x_{i1}) + \ldots + \hat{r}_p(x_{ip}).$$

Then we can write the vector of fitted values $\hat{y} = (\hat{r}(x_1), \ldots \hat{r}(x_n))$ as

$$\hat{y} = \hat{y}^{(1)} + \ldots + \hat{y}^{(p)},$$

where $\hat{y}^{(j)} = (\hat{r}_j(x_{1j}), \ldots \hat{r}_j(x_{nj}))$, for each $j = 1, \ldots p$

- Recall that backfitting cycles through performing univariate smoothing on each dimension, one at a time. Suppose that the linear smoother for dimension j has corresponding smoothing matrix S_j (i.e., this matrix is constructed over the points $x_{1j}, \ldots x_{nj}$). The backfitting updates can then be written as

$$\begin{aligned}
\hat{y}^{(1)} &\leftarrow S_1\Big(y - \sum_{j \neq 1} y^{(j)}\Big) \\
\hat{y}^{(2)} &\leftarrow S_2\Big(y - \sum_{j \neq 2} y^{(j)}\Big) \\
&\ldots \\
\hat{y}^{(n)} &\leftarrow S_n\Big(y - \sum_{j \neq n} y^{(j)}\Big),
\end{aligned}$$

and at convergence, these are all equalities

- This means that there exists linear transformations $R_1, \ldots R_p$ such that the fitted components satisfy

$$\hat{y}^{(j)} = R_j y, \quad j = 1, \ldots p.$$

These fitted values evidently just come from linear smoothers (defined by $R_1, \ldots R_p$), and so we can construct estimates of their variance, confidence intervals, and perform F tests just like in the univariate case