LECTURE 9: DIRECT INFERENCE WITH LINEAR SMOOTHERS

OVERVIEW

For linear regression, we have a well-developed theory for statistical inference. Under normality of regression errors (or large sample sizes), we can easily compute standard errors and make inferences on parameters and mean functions.

What if you go beyond linear regression to, for example, kernel regression and smoothing splines? It turns out that many of the formulas based on standard normal theory (including the F test for whether a regression predicts significantly better than, for example, the global mean) carry over from linear regression to linear smoothers, if one uses the right definition of degrees of freedom, *and* the noise is IID and Gaussian (or sample sizes are large).

Caution: In settings where standard normal theory does not apply, direct inference methods can be misleading; in these cases, the bootstrap can be really useful in getting estimates of error variability (see Lecture 4). Nevertheless, the direct inference tools are computationally more efficient, have close ties to those from linear regression, and are already implemented in R software, so they are worth knowing...

WHAT'S NEXT?

Lecture 10: Logistic regression.

Review of Inference in Linear Regression

The multiple linear regression model is

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon = \beta^\top X + \epsilon \tag{1}$$

where $\beta = (\beta_0, \ldots, \beta_p)^\top$ and $X = (1, X_1, \ldots, X_p)^\top$. The value of the j^{th} covariate for the i^{th} subject is denoted by X_{ij}. Thus

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i.$$

Let

$$\underset{n\times q}{\mathbb{X}} = \begin{pmatrix} 1 & X_{11} & X_{12} & \ldots & X_{1p} \\ 1 & X_{21} & X_{22} & \ldots & X_{2p} \\ \vdots & \vdots & \vdots & \vdots & \\ 1 & X_{n1} & X_{n2} & \ldots & X_{np} \end{pmatrix}$$

Each subject corresponds to one row. **The number of columns of $\mathbb{X}$ corresponds to the number of features plus 1 for the intercept $q = p + 1$.** Now define

$$\vec{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \qquad \vec{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix} \qquad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}. \tag{2}$$

We can then rewrite (1) as

$$\vec{Y} = \mathbb{X}\beta + \vec{\epsilon}. \tag{3}$$

Theorem 1 *If* $(\mathbb{X}^\top\mathbb{X})$ *is invertible, the* least squares estimator *is*

$$\hat{\beta} = (\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top\vec{Y},$$

Inference for the Regression Function

We will review inference for the regression function $r(x)$. (As you learned, inference for the regression coefficients β_j is also possible, but we will focus on the regression function because this is what is relevant for the general regression setting with linear smoothers.)

Fitted values. For linear regression, the fitted values $\hat{\vec{Y}}$ are given by

$$\hat{\vec{Y}} = \mathbb{X}\hat{\beta} = \mathbb{X}(\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top\vec{Y} = H\vec{Y} \tag{4}$$

where $H = \mathbb{X}(\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top$, and the prediction $\hat{r}(x)$ at a new x is given by

$$\begin{aligned}\hat{r}(x) = x^\top\hat{\beta} &= x^\top(\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top\vec{Y} = \left(\mathbb{X}(\mathbb{X}^\top\mathbb{X})^{-1}x\right)^\top\vec{Y} \\ &= \sum_{i=1}^{n}\ell_i(x)Y_i.\end{aligned}$$

Assume that we observe

$$Y_i = \beta^\top x_i + \epsilon_i, \quad \epsilon_i \sim i.i.d.\ N(0,\sigma^2), \quad i = 1,\dots n,$$

where x_i, $i = 1,\dots,n$ are considered fixed. Consider an arbitrary point x.

Question: Do you remember how to derive the following?

(i) an estimate of the variance $\sigma^2 = \mathrm{Var}(\epsilon_i)$,

(ii) the standard error of $\hat{r}(x) = \hat{\beta}^\top x$, and

(iii) a point-wise confidence interval for the regression function $r(x) = \beta^\top x$

(i) Variance σ^2. An unbiased estimator of σ^2 is

Notes:

$$\begin{aligned}\hat{\sigma}^2 &= \frac{\text{RSS}}{\text{residual df}} \\ &= \frac{1}{n-q}\sum_{j=1}^{n}(Y_i - \hat{Y}_i)^2 \\ &= \frac{1}{n-q}\|\vec{Y} - \widehat{\vec{Y}}\|_2^2,\end{aligned}$$

which is the residual sum of squares (RSS) divided by the residual df $= n - q = n - (p+1)$. In addition, if ϵ is normally distributed, then the RSS divided by σ^2 has a Chi-squared distribution,

Notes:

$$\frac{(n-q)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-q}.$$

(See *Elements of Statistical Learning*, section 3.2)

Recall (the χ^2 distribution): If $Z_1, \ldots, Z_k$ are independent standard Normal random variables, then $\sum_{i=1}^{k} Z_i^2 \sim \chi^2_k$. By the CLT, because the chi-squared distribution is the sum of k independent random variables with finite mean and variance, it converges to a normal distribution for large k. For many practical purposes, for $k > 50$ the distribution is sufficiently close to a normal distribution for the difference to be ignored.

(ii) Variance of $\hat{r}(x)$. At an arbitrary fixed point x, we have that

Notes:

$$\begin{aligned}
\hat{r}(x) &= \ell(x)^\top \vec{Y} = x^\top (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top \vec{Y}, \\
\text{Var}(\vec{Y}) &= \sigma^2 I_n, \\
\text{Var}(\hat{r}(x)) &= \ell(x)^\top \text{Var}(\vec{Y}) \ell(x) \\
&= x^\top (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top \text{Var}(Y) \mathbb{X} (\mathbb{X}^\top \mathbb{X})^{-1} x \\
&= \sigma^2 x^\top (\mathbb{X}^\top \mathbb{X})^{-1} x \\
&= \sigma^2 \ell(x)^\top \ell(x).
\end{aligned}$$

By substituting $\hat{\sigma}^2$ for σ^2, we find the estimated variance of $\hat{r}(x)$ is given by

Notes:

$$\hat{\mathsf{se}}^2 = \hat{\sigma}^2 \ell(x)^\top \ell(x) = \hat{\sigma}^2 x^\top (\mathbb{X}^\top \mathbb{X})^{-1} x.$$

Recall (random vectors): Let Y be a random vector. Denote the mean vector by μ and the covariance matrix by $\text{Var}(Y)$ or $\text{Cov}(Y)$ or Σ. If a is a vector then

$$\mathbb{E}(a^\top Y) = a^\top \mu, \quad \text{Var}(a^\top Y) = a^\top \Sigma a.$$

If A is a matrix then

$$\mathbb{E}(AY) = A\mu, \quad \text{Var}(AY) = A\Sigma A^\top.$$

(iii) Point-wise confidence interval for $r(x)$. We have that $\hat{r}(x)$ (after centering and rescaling) follows a t distribution with $n - q$ degrees of freedom; that is

Notes:

$$\frac{\hat{r}(x) - r(x)}{\hat{\mathsf{se}}} \sim t_{n-q}.$$

Suppose that q_1 and q_2 are the $\alpha/2$ and $(1-\alpha/2)$ quantiles of t_{n-q}, respectively. Then, a $(1-\alpha)$ confidence interval for $r(x)$ is given by

Notes:

$$\left[\hat{r}(x) - q_2 \hat{\mathsf{se}}, \hat{r}(x) - q_1 \hat{\mathsf{se}}\right].$$

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):

Notes: For each x,

$$\mathbb{P}\left(\hat{r}(x) - q_2 \hat{\mathsf{se}} \leq r(x) \leq \hat{r}(x) - q_1 \hat{\mathsf{se}}\right) = 1 - \alpha.$$

Often, we will want to construct a confidence interval for the underlying regression function over the *observed* input values, $r(x_i)$, $i = 1, \ldots, n$. From the above, we know that the ith such confidence interval is given by

$$\left[\hat{Y}_i - q_2 \hat{s}(\hat{Y}_i), \hat{Y}_i - q_1 \hat{s}(\hat{Y}_i)\right].$$

Note that the (estimated) variance of $\hat{Y}_i$ is here

$$\hat{s}^2(\hat{Y}_i) = \hat{\sigma}^2 x_i^\top (\mathbb{X}^\top \mathbb{X})^{-1} x_i.$$

Another way of looking at things, in matrix notation:

$$\mathrm{Var}(\hat{\vec{Y}}) = \mathrm{Var}(H\vec{Y}) = \sigma^2 H H^\top = \sigma^2 H,$$

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

Significance Tests Between Fitted Models

We can also test for significance between two fitted nested regression models. Suppose we have two nested sets of variables:

Notes: Let $M_1 \subseteq M_2 \subseteq \{1, \ldots, p\}$ be two nested sets. Here M_1 represents the indices of variables included in model 1, M_2 represents the indices of variables included in model 2, and $\{1, \ldots, p\}$ is the set of all possible variables (the "full" model).

with sizes $p_1 = |M_1|$ and $p_2 = |M_2|$ (counting the intercepts). Let $\hat{\hat{Y}}^{(1)}$ denote the vector of fitted values from the regression on variables in M_1, and $\hat{\hat{Y}}^{(2)}$ from the regression on variables in M_2. Define

$$\text{RSS}_1 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i^{(1)}\right)^2, \quad \text{RSS}_2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i^{(2)}\right)^2,$$

the residual sum of squares from these two regressions. To test the significance of variables in $M_2 \setminus M_1$, i.e., to test the null hypothesis

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

versus the alternative

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

we use the F **statistic**

Notes:

$$F = \frac{(\text{RSS}_1 - \text{RSS}_2)/(p_2 - p_1)}{\text{RSS}_2/(n - p_2)}.$$

If the errors are i.i.d $N(0, \sigma^2)$ or the sample size is large enough, then under the null hypothesis, we have

Notes: Assuming normality,

$$\frac{\text{RSS}}{\sigma^2} \sim \chi^2_{n-p}.$$

When H_0 is true,

$$\frac{\text{RSS}_1 - \text{RSS}_2}{\sigma^2} \sim \chi^2_{p_2 - p_1}.$$

Here the numerator is the deviance and the denominator the degrees of freedom of that deviance.

Putting this together,

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

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 the $(1 - \alpha)$ quantile of $F_{p_2-p_1,n-p_2}$.

Recall (the F distribution): A random variate of the F-distribution with parameters d_1 and d_2 arises as the ratio of two appropriately scaled chi-squared variables: $X = \frac{U_1/d_1}{U_2/d_2}$ where U_1 and U_2 have chi-squared distributions with d_1 and d_2 degrees of freedom respectively, and U_1 and U_2 are independent.

Inference with Linear Smoothers

Recall: An estimator $\hat{r}_n$ of r is a **linear smoother** if, for each x, there exists a vector $\ell(x) = (\ell_1(x), \ldots, \ell_n(x))^\top$ such that

$$\hat{r}(x) = \sum_{i=1}^{n} \ell_i(x) Y_i. \tag{5}$$

(Linear smoothers include e.g. linear regression, k-nearest neighbors regression, kernel regression and smoothing splines.)

Define the vector of **fitted values**

$$\hat{\vec{Y}} = (\hat{r}(x_1), \ldots, \hat{r}(x_n))^\top \tag{6}$$

where $\vec{Y} = (Y_1, \ldots, Y_n)^\top$. It then follows that

$$\hat{\vec{Y}} = S\,\vec{Y} \tag{7}$$

where S is an $n \times n$ matrix whose i^{th} row is $\ell(X_i)^\top$; thus, $S_{ij} = \ell_j(X_i)$. The entries of the i^{th} row show the weights given to each Y_i in forming the estimate $\hat{r}(X_i)$.

The matrix S is called the **smoothing matrix** or the **hat matrix**. The i^{th} row of S is called the **effective kernel** for estimating $r(X_i)$. We define the **effective degrees of freedom** by

$$d = \text{trace}(S). \tag{8}$$

Compare with linear regression where $d = q$. The larger d, the more complex the model. A smaller d yields a smoother regression function.

Linear Smoothers: Inference for the Regression Function

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

$$Y_i = r(x_i) + \epsilon_i, \quad \epsilon_i \sim i.i.d.\ N(0, \sigma^2), \quad i = 1, \ldots, n,$$

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

- Just as in the linear regression case, at an arbitrary point x, the **variance of the fit** $\hat{r}(x) = \ell(x)^\top Y$ is

Notes:

$$\mathrm{Var}\big(\hat{r}(x)\big) = \sigma^2 \ell(x)^\top \ell(x)$$

- **How do we estimate σ^2?** We can now use the estimate

Notes:

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

where $d = \mathrm{df}(\hat{r}) = \mathrm{trace}(S)$, the effective degrees of freedom of the fit $\hat{Y}$.

Note: d replaces q 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** $\hat{r}(x)$

$$\hat{s}^2\big(\hat{r}(x)\big) = \hat{\sigma}^2 \ell(x)^\top \ell(x),$$

and from the same arguments as before, an approximate $(1-\alpha)$ **confidence interval for** $r(x)$, the underlying regression function at a point x, is

$$[\hat{r}(x) - q_2 \hat{s}(\hat{r}(x)), \hat{r}(x) - q_1 \hat{s}(\hat{r}(x))],$$

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

$$\hat{s}^2(\hat{Y}_i) = \hat{\sigma}^2 \ell(x_i)^\top \ell(x_i),$$

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

$$\mathrm{Var}(\hat{\vec{Y}}) = \mathrm{Var}(S\vec{Y}) = \sigma^2 S S^\top,$$

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

The Bias Problem

Confidence bands in regression are not really confidence bands for the true regression function $r(x)$, rather, they are confidence bands for $\bar{r}_n(x) = \mathbb{E}[\hat{r}_n(x)]$, which you can think of a smoothed version of $r(x)$. This is because

Notes:

$$\frac{\hat{r}_n(x) - r(x)}{s_n(x)} = \frac{\hat{r}_n(x) - \bar{r}_n(x)}{s_n(x)} + \frac{\bar{r}_n(x) - r(x)}{s_n(x)}$$
$$= Z_n(x) + \frac{\mathsf{bias}(\hat{r}_n(x))}{\sqrt{\mathrm{Var}(\hat{r}_n(x))}}.$$

Typically, the first term $Z_n(x)$ converges to a standard Normal from which one derives confidence bands. The second term is the bias divided by the standard deviation. In parametric inference, the bias is usually smaller than the standard deviation of the estimator so this term goes to zero as the sample size increases. In nonparametric inference, we have seen that the optimal smoothing corresponds to balancing the bias and the standard deviation. The second term does not vanish even with large sample sizes; the result is that **the confidence interval will not be centered around the true function r due to the smoothing bias $\bar{r}_n(x) - r(x)$.**

Linear Smoothers: Significance Tests Between Fitted Models

Finally, we present an analog of the F test for linear smoothers. 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

Notes:

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

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

Notes:

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

for their respective degrees of freedom, and also

Notes:

$$\text{RSS}_1 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i^{(1)})^2, \quad \text{RSS}_2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i^{(2)})^2,$$

(note that $\text{RSS}_1 \geq RSS_2$, because model 1 is nested in model 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. A linear fit is a special case of a smoothing spline, so model 1 is nested in model 2.

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 $\hat{\hat{Y}}^{(2)}$ over $\hat{\hat{Y}}^{(1)}$ is

Notes:

$$F = \frac{(\text{RSS}_1 - \text{RSS}_2)/(d_2 - d_1)}{\text{RSS}_2/(n - d_2)}.$$

This asks: is the difference in RSS larger than expected with the change in model size, or are we just fitting the noise?

If the errors are i.i.d $N(0, \sigma^2)$ or the the sample size is large enough, then under the null hypothesis, it approximately holds that

Notes:

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

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

R Demo 9.1

Examine the `Wage` data from the ISL book again.
(a) Fit a smoothing spline of wage on age, fixing it at 5 degrees of freedom.
(b) Write a function that returns the smoothing matrix S for splines.
(c) Apply the function to this example and check the result.
(d) Construct an approximate 95% confidence interval for the regression function.