STATS305A - Lecture 10

John Duchi
Scribed by Michael Howes

10/21/21

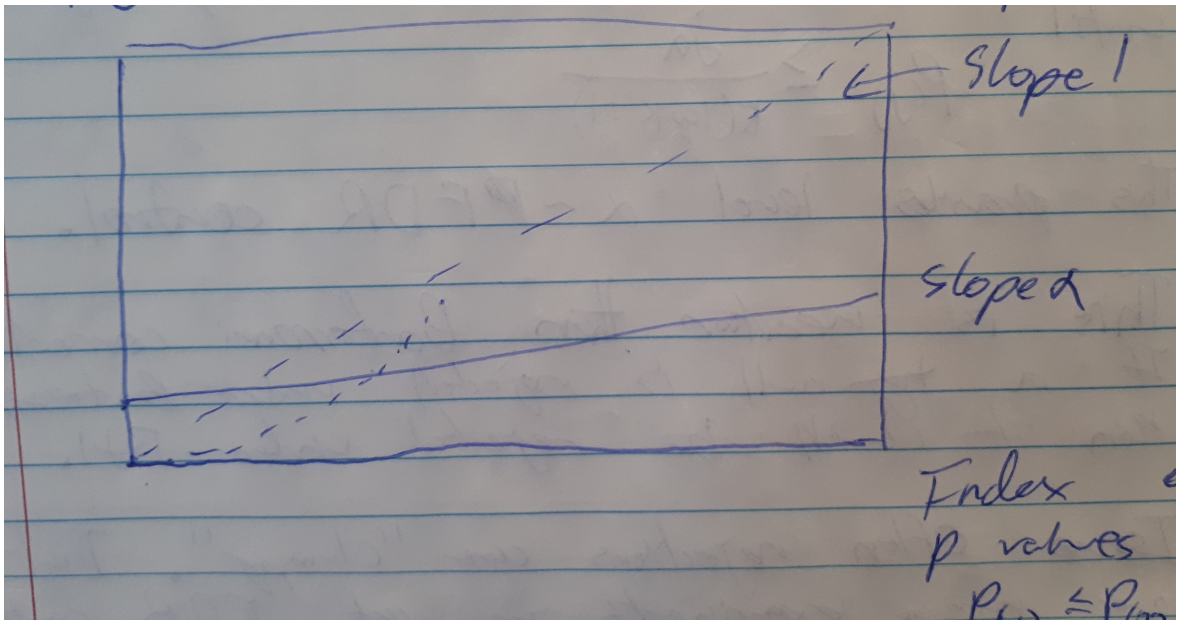
Contents

1 FDR

Setting: We have null hypothesis $H_{0,j}$ for $j = 1, \ldots, k$ and corresponding p-values $p_1, \ldots, p_k$. We define

$$FDR = \mathbb{E}\left[\frac{\#\text{false rejections}}{\#\text{total rejections}}\right].$$

Intuition: p-values should be uniform under H_0. We can thus sort our p-values $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(k)}$. The large p-values should have slope 1 but the smaller p-values which hopefully correspond to false nulls (true discoveries) will have a smaller slope (see picture).



The Benjamini Hochberg prodecure says to reject until $p_{(j)} \geq \frac{j\alpha}{k}$.

Theorem 1. [Benjamini-Hochberg '95] *Let $N_0 = \#$ of true nulls (nulls we shouldn't reject) if the p-values are independent, then BH procedure satisfies*

$$FDR \leq \frac{N_0}{k}\alpha \leq \alpha.$$

If the p-values are dependent, then we always have

$$FDR \leq H(k)\frac{N_0}{k}\alpha \leq H(k)\alpha,$$

where $H(k) = 1 + \frac{1}{2} + \ldots + \frac{1}{k}$ is the k^{th} harmonic number and satisfies $H(k) \leq \log(k) + 1$. Thus if we use the strong criterion reject until

$$p_{(j)} \geq \frac{j\alpha}{\log(k) + 1},$$

then we are guaranteed level α FDR control.

The BH correction is weaker than the Bonferronni correction. If a null is rejected under Bonferroni, then it will still be rejected under BH. Some nulls will be rejected under BH that won't be rejected under Bonferroni.

We still have the issue that often our rejections are "clumpy". That is, we only have control over the average FDP over many experiments. We often see in experiments $\{0\%, 0\%, 0\%, 100\%\}$-false rejections which has an average FDP of 25% but obviously sometimes we have far too many false rejections.

2 Model diagnostics

Motivation: when doing a linear model we want to do as well as possible. Thus we often need to

- Decide when a fit is "good enough" so that the assumptions make sense.
- Choose good features/covariates.

2.1 Leverage scores

Suppose we have the model $Y = X\beta + \varepsilon$. Define $H = X(X^TX)^{-1}X^T$ which is our hat matrix/the orthogonal projection onto the range of X.

We can ask how much does a single example/observation affect our model predictions. That is, what is

$$\frac{\partial \widehat{y}_i}{\partial y_i} = \text{the change in predicted value } \widehat{y}_i \text{ given a change in } y_i.$$

We know that $\widehat{y} = Hy$. Thus if $H = [H_{ij}]_{i,j=1}^n$, then $\widehat{y}_i = \sum_{i=1}^n H_{ij}\widehat{y}_j$ and so $\frac{\partial \widehat{y}_i}{\partial y_i} = H_{ii}$.

Definition 1. The *leverage score* for example i is H_{ii}. (We can interpret this as the self influence of y_i on $\widehat{y}_i$).

Intuition: points with high leverage may actually be problematic as estimates depend strongly on them (although whether or not they are problematic depends on the situation).

Note that

$$H_{ii}(1 - H_{ii}) = \sum_{j \neq i} H_{ij}^2 \geq 0. \tag{1}$$

Thus $H_{ii} \in [0,1]$. Also we have

$$\sum_{i=1}^{n} H_{ii} = \mathrm{tr}(H) = \mathrm{tr}(X(X^TX)^{-1}X^T) = \mathrm{tr}(X^TX(X^TX)^{-1}) = \mathrm{tr}(I_d) = d.$$

Thus if $H_{ii} \geq \frac{2d}{n}$ (or $\frac{3d}{n}$), then the leverage score of i is high (since it is higher than the average d/n). To see equation (1) is true recall that $H^2 = H$ and $H^T = H$. Thus

$$H = H^2 = H^TH = (h_j^T h_i)_{i,j=1}^n,$$

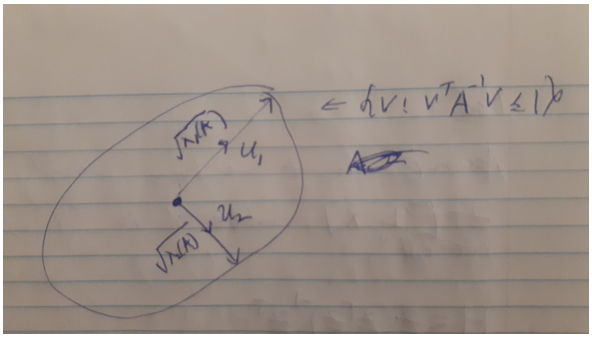
where $H = [h_1, \ldots, h_n]$. Thus

$$H_i i = h_i^T h_i = \sum_{j=1}^{n} H_{ij}^2.$$

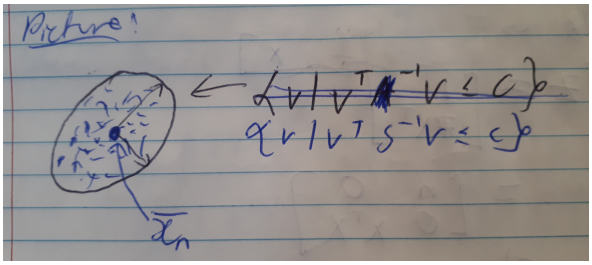
Thus $H_i i - H_{ii}^2 = \sum_{j \neq i} H_{ij}^2$, as claimed.

Definition 2. Given a symmetric matrix $A \succ 0$ (ie $A^T = A$ and A is positive definite). Define $\|x\|_A^2 := x^TAx$ to be the *Mahalanobis norm*. We can then define $D(x,y) = \|x-y\|_A$ to be the *Mahalanobis distance*.

The Mahalanobis norm measures distances scaled by an ellipse. If $A = \sum_{i=1}^{d} \lambda_i(A) u_i u_i^T$, then the set $\{v : v^TA^{-1}v \leq 1\}$ is an ellipse with principal axes u_i with length $\sqrt{\lambda_i(A)}$ (see picture).



Consider now a pile of data points x_i and let $\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$ and define the sample covariance to be $S = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)(x_i - \bar{x}_n)^T$. The matrix S tells us the shape of the "bulk" of our data. See picture:



For $x \in \mathbb{R}^d$, define

$$D^2(x) = (x - \bar{x}_n)^T S^{-1} (x - \bar{x}_n).$$

Thus $D^2(x)$ is a measurement of how far x is from the "bulk" of our data set.

Proposition 1. *If the data has an intercept term, then* $D^2(x_i) = nH_{ii} - 1$.

Proof. Since we have an intercept write $Z = [\mathbf{1}, X]$ where X is "X without the intercept". Define $H = Z(Z^TZ)^{-1}Z^T$. Since our data has an intercept we can assume without loss of generality that $\bar{x}_n = \frac{1}{n}X^T\mathbf{1} = 0$. This implies that

$$\begin{aligned} Z^TZ &= \begin{bmatrix} \mathbf{1}^T \\ X^T \end{bmatrix} [\mathbf{1}, X] \\ &= \begin{bmatrix} n & 0 \\ 0 & X^TX \end{bmatrix}. \end{aligned}$$

Thus

$$(Z^TZ)^{-1} = \begin{bmatrix} \frac{1}{n} & 0 \\ 0 & (X^TX)^{-1} \end{bmatrix},$$

and

$$H = [\mathbf{1}, X] \begin{bmatrix} \frac{1}{n} & 0 \\ 0 & (X^TX)^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{1}^T \\ X^T \end{bmatrix}.$$

Write

$$X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix} \quad \text{and} \quad X^T = [x_1, \ldots, x_n].$$

The diagonal entries of H are thus

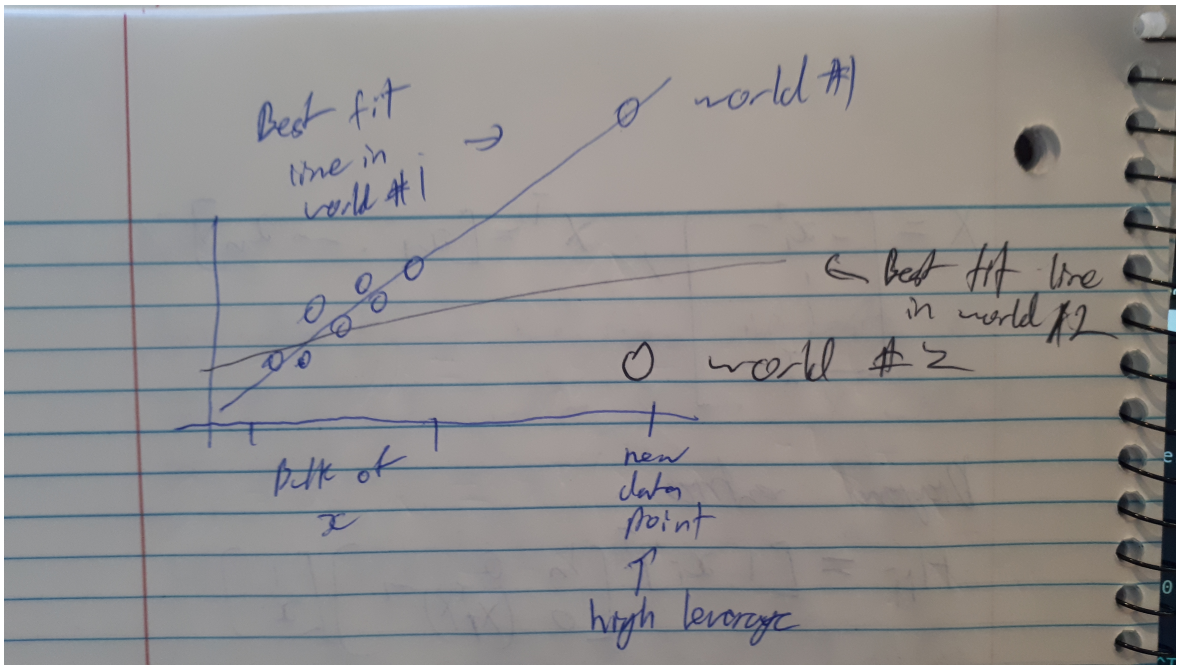
$$H_{ii} = [1, x_i^T] \begin{bmatrix} \frac{1}{n} & 0 \\ 0 & (X^TX)^{-1} \end{bmatrix} \begin{bmatrix} 1 \\ x_i \end{bmatrix} = \frac{1}{n} + x_i^T(X^TX)^{-1}x_i.$$

Note that since $\bar{x}_n = 0$,

$$\begin{aligned} H_{ii} &= \frac{1}{n} + x_i^T(X^TX)^{-1}x_i \\ &= \frac{1}{n}\left(1 + nx_i^T(X^TX)^{-1}x_i\right) \\ &= \frac{1}{n}\left(1 + x_i^T(\frac{1}{n}X^TX)^{-1}x_i\right) \\ &= \frac{1}{n}\left(1 + (x_i - \bar{x}_n)^T\left(\frac{1}{n}X^TX - \bar{x}_n\bar{x}_n^T\right)^{-1}(x_i - \bar{x}_n)\right) \\ &= \frac{1}{n}\left(1 + (x_i - \bar{x}_n)^TS^{-1}(x_i - \bar{x}_n)\right) \\ &= \frac{1}{n}(1 + D^2(x_i)). \end{aligned}$$

Thus $D^2(x_i) = nH_{ii} - 1$. □

Thus high leverage corresponds to the distance of x_i from the bulk of the data. Consider the following picture:



In this picture, the bulk of our data is quite tightly clustered. Consider adding a much larger new data point x_i. This is a point of high leverage by the proposition. We can now consider two worlds corresponding to two different choices of y_i. We have world #1 where the new y_i matches the trend of the rest of the data points. We could also have world #2 where the new data point is far from the trend of the rest of the data.

In both of these world we can make a line of best fit. In world #1 the new point is good and it nails down the slope and suggests our model is good. In model # 2 the new point is not so good. It drastically changes our model. When we have data points like in world # 2 we have to consider two possibilities. It might be the case that the new data point is an outlier due to measurement error. In this case we should consider removing it. If we are very sure of our measurements for the new data, then this will suggest that the model is not correct and we should not just chuck out the new point.

3 Leave one out predictions

Idea: the error in prediction y_i from all data except i should have more "fidelity" to the real world.

Notation 1. Let $H = X(X^TX)^{-1}X^T$,

$$y_{\setminus i} = \begin{bmatrix} y_1 \\ \vdots \\ y_{i-1} \\ y_{i+1} \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^{n-1},$$

and

$$X_{\setminus i} = \begin{bmatrix} x_1^T \\ \vdots \\ x_{i-1}^T \\ x_{i+1}^T \\ \vdots \\ x_n^T \end{bmatrix} \in \mathbb{R}^{(n-1)\times d}.$$

Also define $\widehat{\beta}_{\setminus i} = (X_{\setminus i}^T X_{\setminus i})^{-1} X_{\setminus i}^T y_{\setminus i}$. Thus all of these correspond to the model when we remove y_i. Lastly define $\widehat{y}_{\setminus i} = x_i^T \widehat{\beta}_{\setminus i}$ which is the predicted y-value for x_i without using (x_i, y_i).

Proposition 2. *With the notation as above,* $\widehat{y}_i = H_{ii} y_i + (1 - H_{ii})\widehat{y}_{\setminus i}$.

Proof. Recall the Sherman-Morrison-Woodbury identity (proved on homework)

$$(A - uu^T)^{-1} = A^{-1} + \frac{A^{-1}uu^T A^{-1}}{1 - u^T A^{-1} u},$$

where A is a matrix and u is a vector. Note that

$$X_{\setminus i}^T X_{\setminus i} = X^T X - x_i x_i^T.$$

Set $M = X^T X$, then

$$(X_{\setminus i}^T X_{\setminus i})^{-1} = M^{-1} + \frac{M^{-1} x_i x_i^T M^{-1}}{1 - x_i^T M^{-1} x_i}.$$

Observe that $x_i M^{-1} x_j = H_{ij}$. Now consider $\widehat{y}_i$ and $\widehat{y}_{\setminus i}$. We have

$$\widehat{y}_i = (Hy)_i = H_{ii} y_i + \sum_{j \neq i} H_{ij} y_j.$$

And,

$$\begin{aligned} \widetilde{y}_{\setminus i} &= x_i^T \widehat{\beta}_{\setminus i} \\ &= x_i^T (X_{\setminus i}^T X_{\setminus i})^{-1} X_{\setminus i}^T y_{\setminus i} \\ &= x_i^T \left(M^{-1} + \frac{M^{-1} x_i x_i^T M^{-1}}{1 - x_i^T M^{-1} x_i} \right) X_{\setminus i}^T y_{\setminus i} \\ &= x_i^T M^{-1} X_{\setminus i}^T y_{\setminus i} + \frac{(x_i M^{-1} x_i)(x_i^T M^{-1} X_{\setminus i}^T y_{\setminus i})}{1 - x_i^T M^{-1} x_i}. \end{aligned}$$

Note that since $H_{ij} = x_j^T M^{-1} x_i$, we have $x_i^T M^{-1} x_i = H_{ii}$ and

$$X_{\setminus i}^T (X^T X)^{-1} x_i = [H_{ij}]_{j \neq i}.$$

And so

$$x_i^T (X^T X)^{-1} X_{\setminus i}^T y_{\setminus i} = \sum_{j \neq i} H_{ij} y_j.$$

Subsituting this we have

$$\begin{aligned} \widehat{y}_{\setminus i} &= \sum_{j \neq i} H_{ij} y_j + \frac{H_{ii} \sum_{j \neq i} H_{ij} y_j}{1 - H_{ii}} \\ &= \left(\frac{1 - H_{ii}}{1 - H_{ii}} + \frac{H_{ii}}{1 - H_{ii}} \right) \sum_{j \neq i} H_{ij} y_j \\ &= \frac{1}{1 - H_{ii}} \sum_{j \neq i} H_{ij} y_j. \end{aligned}$$

Thus $(1 - H_{ii})\widehat{y}_{\setminus i} = \sum_{j \neq i} H_{ij} y_j$. Subsituting this into our formula for $\widehat{y}_i$ we can conclude that

$$\widehat{y}_i = H_{ii} y_i + (1 - H_{ii})\widehat{y}_{\setminus i}. \qquad \square$$