

# STATS305A - Lecture 11

John Duchi  
Scribed by Michael Howes

10/26/21

## Contents

<b>1 Announcements</b>	<b>1</b>
<b>2 Leverage</b>	<b>1</b>
<b>3 Residuals, variance and leverage score</b>	<b>2</b>
<b>4 Plots and diagnostics</b>	<b>3</b>
4.1 Residuals vs predicted values . . . . .	3
4.2 Residuals vs features . . . . .	3
4.3 QQ Plots . . . . .	5
4.4 Added variable plot . . . . .	5

## 1 Announcements

- Homework 2 due today.
- Etude 2 due on Thursday.

## 2 Leverage

In the linear model  $Y = X\beta + \varepsilon$ , we define  $H = X(X^T X)^{-1}X^T$  and we define

$$H_{ii} = i^{th} \text{ diagonal entry of } H = i^{th} \text{ leverage score.}$$

We have seen multiple interpretations of leverage score:

- Self influence of example  $i$  on itself.
- $\frac{\partial \hat{y}_i}{\partial y_i} = H_{ii}$ .

We also have the leave one out calculation. We saw that if  $\hat{y}_{\setminus i} = x_i^T \hat{\beta}_{\setminus i}$  where  $\hat{\beta}_{\setminus i} = (X_{\setminus i}^T X_{\setminus i})^{-1} X_{\setminus i}^T Y_{\setminus i}$  and

$$X_{\setminus i} = [x_j^T]_{j \neq i} \in \mathbb{R}^{(n-1) \times d} \quad \text{and} \quad Y_{\setminus i} = [Y_j]_{j \neq i} \in \mathbb{R}^{n-1},$$

then

$$\hat{y}_i = x_i^T \hat{\beta} = (1 - H_{ii}) \hat{y}_{\setminus i} + H_{ii} y_i.$$

Thus (since  $H_{ii} \in [0, 1]$ ),  $\hat{y}_i$  is a convex combination of  $\hat{y}_{\setminus i}$  and  $y_i$ . This also gives us a way to efficiently calculate every  $\hat{y}_{\setminus i}$ . This is something special about the linear model. Note that if we rearrange the above expression we get

$$\hat{y}_{\setminus i} = \frac{1}{1 - H_{ii}} \hat{y}_i - \frac{H_{ii}}{1 - H_{ii}} y_i,$$

Thus

$$[\hat{y}_{\setminus i}]_{i=1}^n = \text{diag}(I - H)^{-1} (\hat{y} - \text{diag}(H)y).$$

Thus in  $O(n)$  times we can do *all* leave one out predictions assuming that we have the matrix  $H$ .

### 3 Residuals, variance and leverage score

Recall  $\hat{\varepsilon} = Y - \hat{Y} = (I - H)Y$ . We always have  $\hat{\varepsilon}^T \hat{Y} = 0$ . This is because

$$\hat{\varepsilon}^T \hat{Y} = \hat{Y}^T (I - H)H\hat{Y} = \hat{Y}^T (H - H)\hat{Y} = 0.$$

The interpretation is that  $\hat{\varepsilon}$  is the noise after removing all “information” in the directions of  $\text{range}(X)$  and  $\hat{Y}$  is the expected value of  $Y$  in the  $X$  directions.

Now suppose we have the distributional assumptions  $Y = X\beta + \varepsilon$  where  $\varepsilon \sim (0, \sigma^2 I)$ . We then have  $\hat{\varepsilon} \sim (0, \sigma^2(I - H))$  and so

$$\text{cov}(\hat{\varepsilon}_i, \hat{\varepsilon}_j) = \begin{cases} \sigma^2(1 - H_{ii}) & \text{if } i = j, \\ -\sigma^2 H_{ij} & \text{else.} \end{cases}$$

And

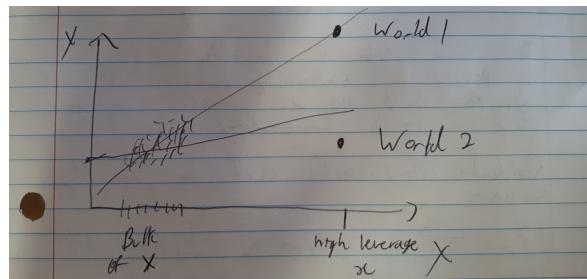
$$\text{Cov}(\hat{y}, \hat{\varepsilon}) = 0.$$

Thus the residuals are typically correlated even when have noise  $\varepsilon \sim (0, \sigma^2 I)$  (in this case we say that the noise is *homoscedastic*). Note also that

$$\text{Var}(\hat{\varepsilon}_i) = \sigma^2(1 - H_{ii}) \leq \sigma^2 = \text{Var}(\varepsilon_i).$$

We call the above decrease in variance, *variance deflation*. The larger the leverage score, the more deflation we have.

Recall the below picture and our intuition that a point with high leverage may be problematic and that the high leverage points are characterized by  $x_i$  being far from the “bulk” of  $X$ . If we have a point of high leverage  $x_i$ , then we could imagine two worlds with different values of  $y_i$ . Since  $x_i$  is high leverage, the point  $\hat{y}_i = x_i^T \hat{\beta}$  changes a lot as  $y_i$  changes. This means that our model is more sensitive to changes in  $y_i$  than changes in  $y$  at points with low leverage.



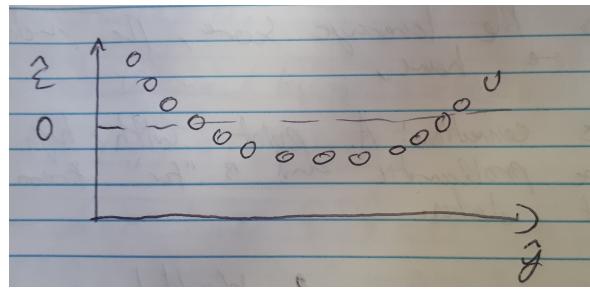
## 4 Plots and diagnostics

How do we decide if our linear models are “good enough”? We have a number of possible desiderata. Some plots which are good sanity checks

- Residuals vs fitted values.
- Residuals vs individual features ie columns in  $X$ .
- Residuals vs quantiles of Gaussians called a QQ or quantile quantile plot.

### 4.1 Residuals vs predicted values

One is that the residual  $\hat{\varepsilon}$  should be *really* uncorrelated with the predicted values  $\hat{y}$ . Consider the below plot:



The above data is uncorrelated but there is clear a relationship between  $\hat{\varepsilon}$  and  $\hat{y}$ . You should always plot your residuals against your predicted values to see they are *really* uncorrelated. In the above example we can conclude that the linear model is probably false and we really want to include a term of the form  $(x^T \beta)^2$  in our model. Unfortunately this make our model non-linear and there's not much we can do.

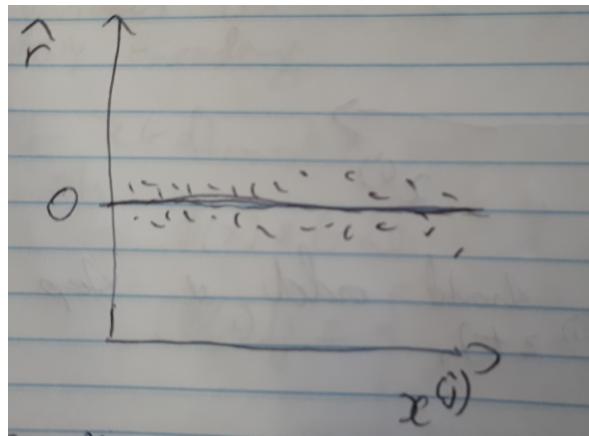
The standardized residuals should be roughly Gaussian. That is define

$$\hat{r}_i = \frac{\hat{\varepsilon}_i}{S_n \sqrt{1 - H_{ii}}},$$

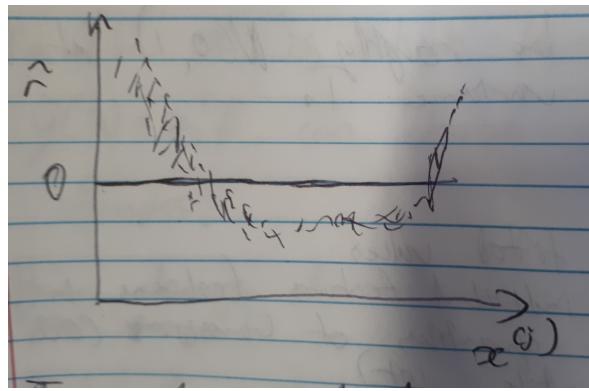
where  $S_n^2 = \frac{1}{n-d} \|X\beta - Y\|_2^2 = \frac{1}{n-d} \|\hat{\varepsilon}\|_2^2$ . These normalized residuals should be roughly  $N(0, 1)$  at least they should have variance 1.

### 4.2 Residuals vs features

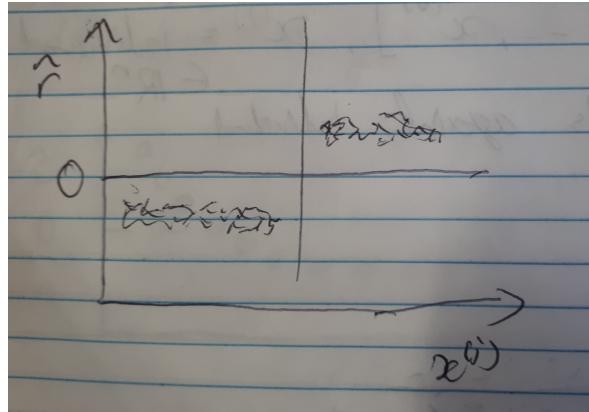
Suppose  $X = [x^{(1)}, \dots, x^{(d)}]$  so  $x^{(j)} \in \mathbb{R}^n$  is an individual feature. We could plot the residuals against the individual features. We might get a plot like this:



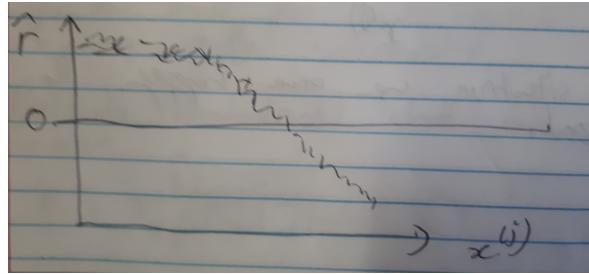
In this situation we are happy and the relationship between the residuals and the feature looks good. We might also have:



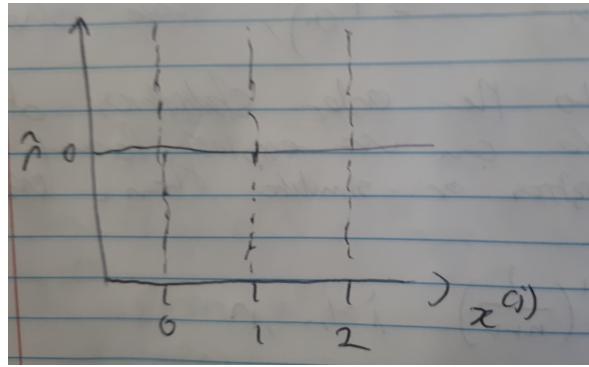
This suggest we should add a polynomial feature like  $(x^{(j)})^2$  to our model. We might also have



In which case we should add a feature which is a step function like  $\mathbf{1}(x^{(j)} > t)$  where  $\mathbf{1}(A)$  is the indicator function of  $A$ . We might have



In this case we should add  $(x^{(j)} - t)_+$ , the positive part of  $x^{(j)} - t$ . Lastly we might have



Which means that  $x^{(j)}$  is a categorical variable and we should not include it “raw” in our model. We should re-encode the variable. There are many names for this re-encoding such as one-hot encoding, dummy variables, factors,  $\{0, 1\}$ -encoding. These all refer to the same process:

If  $x \in \{1, \dots, k\}$ , we replace  $x$  with  $\phi(x) \in \{0, 1\}^k$  given by

$$[\phi(x)]_j = \begin{cases} 1 & \text{if } x = j, \\ 0 & \text{else.} \end{cases}$$

This is similar to the encoding used for  $k$ -groups ANOVA.

### 4.3 QQ Plots

Idea: We can compare the quantiles of  $\hat{r}_i = \frac{\hat{\varepsilon}_i}{s_n \sqrt{1 - H_{ii}}}$  to the quantiles of a standard normal. First we sort the standardized residuals  $\hat{r}_{(1)} \leq \hat{r}_{(2)} \leq \dots \leq \hat{r}_{(n)}$ , these should be similar to the order statistics of a Gaussian. We could use analytic formulas for these order statistics or simulate them. In practice we can also use

$$v_{(i)} = \Phi^{-1} \left( \frac{i}{n+1} \right), \quad \text{for } i = 1, \dots, n.$$

We would then expect  $v_{(i)} \approx \hat{r}_{(i)}$ .

### 4.4 Added variable plot

These give us a sense of how much a variable adds to a model when “adjusting” for other variables.

The mathematical insight: Suppose  $X = [x^{(1)}, \dots, x^{(d)}]$ , then the component of  $\hat{\beta}_j$  in ordinary least squares is equal to the regression of  $y$  on  $x^{(j)}$  adjusting for other variables.

By adjusting we mean we first do regression on  $x^{(j)}$  using all the other variables then we subtract this and take  $\hat{x}^{(j)}$  to be what's left. Then we use  $\hat{x}^{(j)}$  to do regression on  $y$ .

Let's prove this for  $\hat{\beta}_d$ . Suppose  $X = QR$  where  $Q \in \mathbb{R}^{n \times d}$  satisfies  $Q^T Q = I_d$  and  $R \in \mathbb{R}^{d \times d}$  is invertible and upper triangular. Write  $Q = [q_1, \dots, q_d]$ . We know from Gram-Schmidt that if

$$\hat{x}^{(d)} = x^{(d)} - \sum_{j=1}^{d-1} q_j \langle q_j, x^{(d)} \rangle = (I_n - Q_{\setminus d} Q_{\setminus d}^T) x^{(d)},$$

then  $\hat{x}^{(d)}$  is orthogonal to  $q_j$  for  $j < d$ . Thus we can think of  $\hat{x}^{(d)}$  as what is left from  $x^{(d)}$  after taking away the information in the other variables. We can think of  $\hat{x}^{(d)}$  as  $x^{(d)}$  with the “best prediction” of  $x^{(d)}$  from  $x^{(1)}, \dots, x^{(d-1)}$  removed. We say  $\hat{x}^{(d)}$  is  $x^{(d)}$  *adjusted for*  $x^{(1)}, \dots, x^{(d-1)}$ .

Recall that  $q_d = \frac{\hat{x}^{(d)}}{\|\hat{x}^{(d)}\|_2}$  and  $R_{dd} = \|\hat{x}^{(d)}\|_2$ . Let's solve the normal equations. Note that

$$\begin{aligned} X^T X \beta &= X^T Y \iff R^T Q^T Q R \beta = R^T Q^T Y \\ &\iff R \beta = Q^T Y. \end{aligned}$$

Thus if we look at the  $d^{th}$  component, we see that  $\hat{\beta}_d$  must solve  $R_{dd} \beta_d = q_d^T y$ . Thus

$$\hat{\beta}_d = \frac{(\hat{x}^{(d)})^T y}{\|\hat{x}^{(d)}\|_2^2}.$$

Now consider one dimensional linear regression of  $y$  on  $z \in \mathbb{R}$ . We wish to find

$$\min_{b \in \mathbb{R}} \sum_{i=1}^n (y_i - z_i b)^2.$$

The minimizer  $b^*$  satisfies  $\sum_{i=1}^n z_i y_i = (\sum_{i=1}^n z_i^2) b^*$  and so  $b^* = \frac{z^T y}{\|z\|_2^2}$ .

Upshot:  $\hat{\beta}_j$  is the regression of  $y$  onto  $x^{(j)}$  adjusting for  $x^{(k)}$  for all  $k \neq j$ . This gives us an idea of something to plot. If we wish to see if the variable  $j$  should be included in our model we can plot  $y$  adjusted for  $x^{(k)}$  for  $k \neq j$  against  $\hat{x}^{(j)}$ . We can then look for a strong linear relationship in the plot. If there is a strong relationship we should include  $x^{(j)}$ .