Week 2: Linear Regression and Quiz

Advanced Linear Models Reading Group

March, 13, 2017

1 Connection With Linear Regression

We want to find the best fit line through a set of n datapoints (x_i, y_i). Let $\mathbf{X} = \begin{bmatrix} x_i \\ \vdots \\ x_n \end{bmatrix}$ and $\mathbf{Y} = \begin{bmatrix} y_i \\ \vdots \\ y_n \end{bmatrix}$. We want to find $\hat{\beta}_0$ and $\hat{\beta}_1$ such that

$$\|\mathbf{Y} - (\beta_0 \mathbf{1}_n + \beta_1 \mathbf{X})\|^2 \tag{1}$$

is minimized (i.e. find the line of best fit to the data, $y = \hat{\beta}_0 + \hat{\beta}_1 x$).

1. Fix β_1, Find $\hat{\beta}_0(\beta_1)$: If we fix β_1 so that $\mathbf{Z} = \mathbf{Y} - \beta_1 \mathbf{X}$, then our problem becomes minimizing $\|\mathbf{Z} - \beta_0 \mathbf{1}_n\|^2$. This is simply regression with a constant, for which the solution is

$$\hat{\beta}_0(\beta_1) = \bar{\mathbf{Z}} = \bar{\mathbf{Y}} - \beta_1 \bar{\mathbf{X}}. \tag{2}$$

2. Plug in $\hat{\beta}_0$, It's the same as just centering and then regressing through origin: Plugging 2 into 1, we are left with

$$\|\mathbf{Y} - \bar{\mathbf{Y}} - \beta_1 \bar{\mathbf{X}} - \beta_1 \mathbf{X}\|^2 = \|(\mathbf{Y} - \bar{\mathbf{Y}}) - \beta_1 (\mathbf{X} - \bar{\mathbf{X}})\|^2. \tag{3}$$

   This is the same thing as just centing the data before regressing though the origin as discussed previously. Thus $\hat{\beta}_1 = \hat{\rho}_{XY} \frac{\hat{\sigma}_Y}{\hat{\sigma}_X}$ and $\hat{\beta}_0 = \bar{\mathbf{Y}} - \left(\hat{\rho}_{XY} \frac{\hat{\sigma}_Y}{\hat{\sigma}_X}\right) \bar{\mathbf{X}}$.

2 Residuals

Define the residuals of our regression as $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$ where $\hat{\mathbf{Y}} = \hat{\beta}_0 + \hat{\beta}_1 \mathbf{X}$. We show that for LSR with an intercept the sum of the residuals is zero:

$$\sum_n (y_i - \hat{y}_i) = \sum_n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = \sum_n (y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i) = \sum_n (y_i - \bar{y}) + \hat{\beta}_1 \sum_n (x_i - \bar{x}) = 0. \tag{4}$$

3 Quiz

1. *Let $\tilde{X}$ and $\tilde{Y}$ be mean-centered versions of the vectors X and Y. Which of the following are the result of regression through the origin of $\tilde{X}$ and $\tilde{Y}$?*

   (a) $\hat{\rho}_{XY}\frac{\hat{\sigma}_Y}{\hat{\sigma}_X}$ (Yes)

   (b) $\frac{<X,Y>}{<X,X>}$ (No)

   (c) $\frac{<\tilde{X},\tilde{Y}>}{<\tilde{X},\tilde{X}>}$ (Yes)

2. *Let $\tilde{X}$ and $\tilde{Y}$ be mean-centered versions of the vectors X and Y that have also been scaled by their standard deviations ($\sigma_{\tilde{X}} = \sigma_{\tilde{Y}} = 1$). Regression through the origin will give the same slope regardless of whether $\tilde{X}$ is the predictor and $\tilde{Y}$ the outcome or vice versa?* True, notice that in either case $\hat{\beta} = \hat{\rho}_{XY}$.

3. *With regression through the origin must the residuals always sum to zero?* False. This is true for linear regression with an intercept so that

$$\|e\| = \|\mathbf{Y} - \hat{\beta}_0 \mathbf{1}_n - \hat{\beta}_1 \mathbf{X}\| = \sum_n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0. \tag{5}$$

   It follows that

$$\sum_n (y_i - \hat{\beta}_1 x_i) = n\hat{\beta}_0. \tag{6}$$

4. *The inner product of the residuals with the predictor is equal to zero?* True. Observe that

$$< e, \mathbf{X} >= \left(\mathbf{Y} - \mathbf{X}\frac{< \mathbf{X}, \mathbf{Y} >}{< \mathbf{X}, \mathbf{X} >}\right)^T \mathbf{X} = \mathbf{Y}^T\mathbf{X} - \left(\frac{< \mathbf{X}, \mathbf{Y} >}{< \mathbf{X}, \mathbf{X} >}\right)\mathbf{X}^T\mathbf{X} =< \mathbf{Y}, \mathbf{X} > - < \mathbf{Y}, \mathbf{X} >= 0. \tag{7}$$

   A more geometric way to think about this is that $\mathbf{e}$ is orthogonal to $\hat{\mathbf{Y}} = \hat{\beta}_1 \mathbf{X}$.