Math 390.4
Tenth Theoretical Lecture

Joseph Peltroche

3/7/18

Random variables X,Y are said to be dependent if knowing the value of affects the distribution of another. Mathematically:

$$\mathbb{P}(Y|X = x) \neq \mathbb{P}(Y)$$

This follows the same idea of "association": if knowing a predictor x's value allows to know "some thing" about y, then x,y are said to be associative.

Covariance of two random variables X,Y are defined to be

$$\text{Cov}[X, Y] = E[(X - \mu_x)(Y - \mu_y)] \in \mathbb{R}$$

estimated by

$$S_{xy} = \frac{1}{n-1} \sum_{i}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

To better interpret covariance, we can take a look at the correlation ρ, which is a scaled covariance

$$\rho = \text{Cor}[X, Y] = \frac{\text{Cov}[X, Y]}{SE[X]SE[Y]} \in [-1, 1]$$

which is within an intrepretable space $[-1, 1]$, estimated by $r = \frac{S_{xy}}{S_x S_y} \in [-1, 1]$

x,y are "positively correlated" if $r > 0$ meaning $x \uparrow \Rightarrow y \uparrow$

x,y are "negatively correlated" if $r > 0$ meaning $x \uparrow \Rightarrow y \downarrow$

x,y are "not correlated" if $r = 0$ meaning $x \uparrow \Rightarrow y$ (unchanged)

Here, correlation has some relation to association. correlation $\in$ association.

Recall for the linear regression of $\mathcal{Y} \in \mathbb{R}$ and $p = 1$, $\mathcal{H} = \{w_0 + w_1 x_1 : w_0, w_1 \in \mathbb{R}\}$. If $p = 2$, then similarly, our $\mathcal{H}$ takes the form $\mathcal{H} = \{w_0 + w_1 x_1 + w_2 x_2 : w_0, w_1, w_2 \in \mathbb{R}\}$. Using the method of Least squares, we would take partials of the SSE and set them equal to zero to find the w_0, w_1, w_2 that produce a minimum.

$$SSE = \sum_i^n (y_i - \hat{y}_i)^2 = \sum_i^n (y_i - (w_0 + w_1 x_1 + w_2 x_2))^2$$

$$\frac{\partial}{\partial w_0} SSE = 0 \qquad \frac{\partial}{\partial w_1} SSE = 0 \qquad \frac{\partial}{\partial w_2} SSE = 0$$

This would follow for the general p case, and there would be $p + 1$ partials to take.

For the general p case, our X matrix, now with a first column of 1's, would belong to $\mathbb{R}^{n \times (p+1)}$ and $\vec{w} \in \mathbb{R}^{p+1}$. So,

$$X\vec{w} = \begin{bmatrix} w_0 + w_1 x_{11} + \ldots + w_{p+1} x_{1(p+1)} \\ \vdots \\ w_0 + w_1 x_{n1} + \ldots + w_{p+1} x_{n(p+1)} \end{bmatrix} = \begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \vec{\hat{y}} \in \mathbb{R}^{n \times 1}$$

$$\boxed{\vec{\hat{y}} = X\vec{w}}$$

We can extend this idea to our SSE:

$$\begin{aligned} SSE = \sum_i^n (y_i - \hat{y}_i)^2 &= (\vec{y} - \vec{\hat{y}})^T (\vec{y} - \vec{\hat{y}}) && (\vec{v}^T \vec{v} = ||\vec{v}||^2) \\ &= (\vec{y}^T - \vec{\hat{y}}^T)(\vec{y} - \vec{\hat{y}}) \\ &= \vec{y}^T \vec{y} - \vec{y}^T \vec{\hat{y}} - \vec{\hat{y}}^T \vec{y} + \vec{\hat{y}}^T \vec{\hat{y}} \\ &= \vec{y}^T \vec{y} - 2\vec{\hat{y}}^T \vec{y} + \vec{\hat{y}}^T \vec{\hat{y}} && (a^T b = b^T a) \\ &= \vec{y}^T \vec{y} - 2(X\vec{w})^T \vec{y} + (X\vec{w})^T (X\vec{w}) \\ &= \vec{y}^T \vec{y} - 2\vec{w}^T X^T \vec{y} + \vec{w}^T X^T X \vec{w} && ((ab)^T = b^T a^T) \end{aligned}$$

Minimizing this function will be analogous to the OLS method, but replacing scalars for vectors. We want to take derivative in such manner:

$$\frac{\partial}{\partial \vec{w}} SSE = \begin{bmatrix} \frac{\partial}{\partial w_0} SSE \\ \frac{\partial}{\partial w_1} SSE \\ \vdots \\ \frac{\partial}{\partial w_p} SSE \end{bmatrix} = \vec{0}_{p+1}$$

To solidify the concept, here are some examples:

e.g. for a constant $a \in \mathbb{R}$, $\vec{c} \in \mathbb{R}^n$

$$\frac{\partial}{\partial \vec{c}} [a] = \begin{bmatrix} \frac{\partial}{\partial c_1} a \\ \frac{\partial}{\partial c_2} a \\ \vdots \\ \frac{\partial}{\partial c_n} a \end{bmatrix} = \vec{0}_{p+1}$$

e.g. for vectors $\vec{a}, \vec{c} \in \mathbb{R}^n$

$$\frac{\partial}{\partial \vec{c}} [\vec{c}^T \vec{a}] = \begin{bmatrix} \frac{\partial}{\partial c_1} [c_1 a_1 + \ldots + a_n c_n] \\ \frac{\partial}{\partial c_2} [c_1 a_1 + \ldots + a_n c_n] \\ \vdots \\ \frac{\partial}{\partial c_n} [c_1 a_1 + \ldots + a_n c_n] \end{bmatrix} = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} = \vec{a}$$

e.g. for $A \in \mathbb{R}^{n \times n}, \vec{c} \in \mathbb{R}^n$ where A is symmetric.

$$\frac{\partial}{\partial \vec{c}} [\underbrace{\vec{c}^T A \vec{c}}_{\text{quadratic form}}] = \frac{\partial}{\partial \vec{c}} [\vec{c}^T (A\vec{c})] \tag{10.2}$$

$$A\vec{c} = \begin{bmatrix} a_{11}c_1 + a_{12}c_2 + \ldots + a_{1n}c_n \\ \vdots \\ a_{n1}c_1 + a_{n2}c_2 + \ldots + a_{nn}c_n \end{bmatrix} \in \mathbb{R}^{n \times 1}$$

$$\vec{c}^T (A\vec{c}) = c_1(a_{11}c_1 + a_{12}c_2 + \ldots + a_{1n}c_n) + c_2(a_{21}c_1 + a_{22}c_2 + \ldots + a_{2n}c_n) + \ldots$$
$$\ldots + c_n(a_{n1}c_1 + a_{n2}c_2 + \ldots + a_{nn}c_n)$$

So, taking a look at just one partial derivative

$$\frac{\partial}{\partial c_1} [\vec{c}^T (A\vec{c})] = 2c_1 a_{11} + c_2 a_{12} + c_3 a_{13} + \ldots + c_n a_{1n} + c_2 a_{21} + c_3 a_{31} + \ldots + c_n a_{n1}$$

$$= 2(c_1 a_{11} + c_2 a_{12} + \ldots + c_n a_{1n}) \qquad \text{(By symmetry of } A\text{)}$$

A similar outcome results for all proceeding partial derivaitves

$$\frac{\partial}{\partial c_n}[\vec{c}^T(A\vec{c})] = 2(c_1 a_{1n} + c_2 a_{2n} + \ldots + a_{nn})$$

Every row is a multiple of the dot product between a row of A and $\vec{c}$, hence, plugging back in for (10.1)

$$\frac{\partial}{\partial \vec{c}}[\vec{c}^T(A\vec{c})] = 2A\vec{c}$$

Now we can apply this to our vector and matrix expression of SSE

$$\begin{aligned}
&\frac{\partial}{\partial \vec{w}}[\vec{y}^T\vec{y} - 2\vec{w}^T X^T \vec{y} + \vec{w}^T(X^T X)\vec{w}] \\
&= \vec{0}_{p+1} - \cancel{2}X^T\vec{y} + \cancel{2}X^T X\vec{w} = \vec{0}_{p+1} && \text{(set to } \vec{0}_{p+1} \text{ to find extrema)} \\
&\Rightarrow (X^T X)^{-1} X^T X \vec{w} = (X^T X)^{-1} X^T \vec{y}
\end{aligned}$$

$$\boxed{\vec{b} = (X^T X)^{-1} X^T \vec{y}}$$

Note, the above derivations assumed $X^T X$ was a symmetric matrix (fairly easy to prove) and that it is invertible. According to equivalent statements in first semester linear algebra, this would mean $X^T X$ would have a **full rank** of $p+1$. In other words, all the columns are *linearly independent*.

This would mean the $rank(X) = p+1$

Proof. We proceed by contradiction. Assume $rank(X^T X) = p+1$ and $rank(X) < p+1$. Then there exists a vector $\vec{u} \neq \vec{0} \in \mathbb{R}^{p+1}$ such that

$$X\vec{u} = \vec{0}_{p+1}$$

Similarly, we can matrix multiply this to $X^T X$

$$X^T X\vec{u} = X^T(X\vec{u}) = X^T\vec{0}_{p+1} = \vec{0}_{p+1}$$

which violates one of the equivalent statements and is thereby a contradiction. Thus, $X^T X$ does not have a full rank. □

Our expression for $\vec{\hat{y}}$ is now

$$\vec{\hat{y}} = X\vec{b} = \underbrace{X(X^T X)^{-1} X^T}_{\text{“hat matrix”}}\vec{y} = \vec{H}\vec{y} \qquad \text{(plugging in for b)}$$

Where $\vec{H}$ is our “hat matrix”. This matrix sets up the linearly independent columns that will create the linear combinations for each our $\hat{y}_i$'s.