Math 342W Lecture 23

If you see correlation, there will be causation somewhere

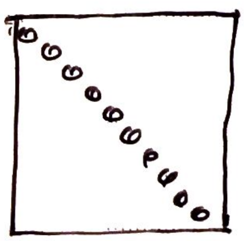
Spurious Correlation: Concluding x and y are correlated when they're not

If $\hat{y} = b_0 + b_1 x_1 + \dots + b_p x_p$, how can we interperet b_1

Wrong: Holding the rest of the features constant, a change of 1 in x_1 results in b_1 change in y's mean.

Correct: When comparing two mutually observed observation (A) and (B) are sampled in the same way as observations in the training set where (A) has a x_1 value one unit larger than the x_1 of (B), and share same values $x_2, \dots, x_p$ then (A) is predicted to have a response y that differs by b_1 units on average from response of (B), assuming linear model is true.

$X =$  p, n, $p > n$

$\Rightarrow X^T X$ 

$\Rightarrow \vec{b}_{ridge} = (X^T X + \lambda I)^{-1} X^T \vec{y}$

$X^T X$ not invertible now, $\vec{b}_{OLS}$ DNE...

Essentially Cantor's Diagolization plus a little "shift" to add more n.

Ridge Regression & Lasso Regression

Diagonalization: $X^TX = VDV^{-1}$

$(p+1)\times(p+1)$

Cols are $p+1$ Eigen Vectors, $\text{rank}[V] = p+1$

$\begin{bmatrix} \lambda_1 & & 0 \\ & \lambda_2 & \\ 0 & & \lambda_n \\ & & & 0 \end{bmatrix}$ Diagonal with Eigen Values

$$\vec{b}_{ridge} = \underbrace{(X^TX + \lambda I)^{-1}}X^T\vec{y}$$

$$(VDV^{-1}+\lambda I)^{-1} = (VDV^{-1} + \lambda I VV^{-1})^{-1} = (VDV^{-1} + V(\lambda I)V^{-1})^{-1}$$

$$= (V(D+\lambda I)V^{-1})^{-1} = V(D+\lambda I)^{-1}V^{-1} \Rightarrow \begin{bmatrix} \lambda_1+\lambda & & & & 0 \\ & \lambda_2+\lambda & & & \\ & & \ddots & & \\ 0 & & & \lambda_n+\lambda & \\ & & & & \ddots & \\ & & & & & \lambda \end{bmatrix}$$

Now we have shifted λ's up to n on the diagonal and 0+λ's up to $p+1$ so now we are full rank.

$\lambda \in \mathbb{R} > 0$

Consider $\mathcal{A}$: $\vec{b}_{ridge} = \text{argmin}\{SSE + \lambda\|\vec{\omega}\|_2^2\}$

Consider $\mathcal{A}$: $\vec{b}_{lasso} = \text{argmin}\{SSE + \lambda\|\vec{\omega}\|_1\}$

$\Rightarrow$ Regularization $(\vec{b} \to \vec{0})$

Roughly speaking, Ridge is used for prediction and Lasso is used for feature selection.

Consider $\mathcal{A}$: $\vec{b}_{elastic} = \text{argmin}\{SSE + \lambda(\alpha\|\vec{\omega}\|_1 + (1-\alpha)\|\vec{\omega}\|_2^2\}$