Спецкурс "Регрессия"

Шкляев А.В.

15 января 2019 г.

0.1 Коэффициент корреляции и частный коэффициент корреляции

Линейная регрессия тесно связана с понятием корреляции. Предположим, что есть случайный вектор X, Y. Тогда

$$\cos\alpha = \frac{\langle X, Y\rangle}{||X||||Y||}$$

представляет собой косинус угла между X и Y, рассматриваемыми как элементами пространства $L^2(P)$. Предположим, что мы хотим убрать влияние константных случайных величин и рассмотреть величины $\widetilde{X} = X - d_X$, $\widetilde{Y} = Y - d_Y$, где d_X, d_Y некоторые константы. Выберем d_X и d_Y так, чтобы $\widetilde{X}$ ортогонально пространству констант c, то есть $\langle X - d_X, c\rangle = \mathbf{E}c(X - d_X) = 0$. Следовательно, $d_X = \mathbf{E}X$. Аналогично $d_Y = \mathbf{E}Y$. Косинус угла между $\widetilde{X}$ и $\widetilde{Y}$

$$\mathbf{corr}(X, Y) = \frac{\langle X - \mathbf{E}X, Y - \mathbf{E}Y\rangle}{||X - \mathbf{E}X||||Y - \mathbf{E}Y||} = \frac{\mathbf{cov}(X, Y)}{\sqrt{\mathbf{D}X\mathbf{D}Y}}$$

называют коэффициентом корреляции между X, Y. Аналогичным образом можно определить частный коэффициент корреляции

$$\mathbf{corr}(X, Y|M),$$

где M — подпространство L^2, содержащее все линейные функции $a_1Z_1 + ... + a_mZ_m + a$ от заданных величин $Z_1, ..., Z_m$. Выберем такие случайные величины E_X, E_Y в M, что $X - E_X$ и $Y - E_Y$ ортогональны M и положим

$$\mathbf{corr}(X, Y|M) = \mathbf{corr}(X, Y|Z_1, ..., Z_m) = \frac{\langle X - E_X, Y - E_Y\rangle}{||X - E_X||||Y - E_Y||} = \frac{\mathbf{E}(X - E_X)(Y - E_Y)}{\sqrt{\mathbf{E}(X - E_X)^2\mathbf{E}(Y - E_Y)^2}} = \mathbf{corr}(X - E_X, Y - E_Y),$$

где в последнем равенстве мы воспользовались тем, что $\mathbf{E}(X - E_X) = \mathbf{E}(Y - E_Y) = 0$ т.к. $X - E_X$, $Y - E_Y$ ортогональны константе a. При этом $E_X = a_{X,1}Z_1 + .. + a_{X,m}Z_m + a_X$, $E_Y = a_{Y,1}Z_1 + ... + a_{Y,m}Z_m + a_Y$, коэффициенты могут быть найдены из соотношений $X - E_X \perp M$, $Y - E_Y \perp M$:

$$\mathbf{E}XZ_i = \langle \vec{a_{X,\cdot}}, \mathbf{E}Z_i\vec{Z}\rangle + a_X\mathbf{E}Z_i,\ \mathbf{E}X = \langle \vec{a_{X,\cdot}}, \mathbf{E}\vec{Z}\rangle + a_X, \mathbf{E}YZ_i = \langle \vec{a_{Y,\cdot}}, \mathbf{E}Z_i\vec{Z}\rangle + a_Y\mathbf{E}Z_i,\ \mathbf{E}Y = \langle \vec{a_{Y,\cdot}}, \mathbf{E}\vec{Z}\rangle + a_Y.$$

В частном случае при $m = 1$

$$\mathbf{E}XZ = a_{X,1}\mathbf{E}Z^2 + a_X\mathbf{E}Z,\ \mathbf{E}X = a_{X,1}\mathbf{E}Z + a_X,$$

откуда $a_{X,1} = \mathbf{cov}(X,Z)/\mathbf{D}Z$. Аналогичным образом находится $a_{Y,1}$, откуда

$$\mathbf{corr}(X,Y|Z) = \frac{\mathbf{cov}(X - a_{X,1}Z, Y - a_{Y,1}Z)}{\sqrt{\mathbf{D}(X - a_{X,1}Z)\mathbf{D}(Y - a_{Y,1}Z)}} = \frac{\mathbf{cov}(X,Y) - \frac{\mathbf{cov}(Z,Y)\,\mathbf{cov}(Z,X)}{\mathbf{D}Z}}{\sqrt{\left(\mathbf{D}X - \frac{\mathbf{cov}(X,Z)^2}{\mathbf{D}Z}\right)\left(\mathbf{D}Y - \frac{\mathbf{cov}(Y,Z)^2}{\mathbf{D}Z}\right)}} =$$

$$\frac{\mathbf{corr}(X,Y) - \mathbf{corr}(Y,Z)\,\mathbf{corr}(X,Z)}{\sqrt{(1 - \mathbf{corr}(X,Z))\,(1 - \mathbf{corr}(Y,Z))}}$$

Пример 1. Пусть $X = Z + X_1$, $Y = 2Z + X_2$, где Z, X_1, X_2 независимы. Тогда

$$\mathbf{corr}(X,Y) = \frac{\mathbf{cov}(X,Y)}{\sqrt{\mathbf{D}X\mathbf{D}Y}} = \frac{2\mathbf{D}Z}{\sqrt{(\mathbf{D}Z + \mathbf{D}X_1)(4\mathbf{D}Z + \mathbf{D}X_2)}}.$$

При этом в числителе $\mathbf{corr}(X,Y|Z)$ расположена величина

$$\mathbf{corr}(X,Z) = \frac{2\mathbf{D}Z}{\sqrt{(\mathbf{D}Z + \mathbf{D}X_1)(4\mathbf{D}Z + \mathbf{D}X_2)}} - \frac{\mathbf{D}Z}{\sqrt{(\mathbf{D}Z + \mathbf{D}X_1)\mathbf{D}Z}}\frac{2\mathbf{D}Z}{\sqrt{(4\mathbf{D}Z + \mathbf{D}X_1)\mathbf{D}Z}} = 0.$$

Таким образом, X, Y некоррелированы при условии Z и исключенный коэффициент корреляции устранил влияние Z.

В общем случае можно выражать частную корреляцию поступательно

$$\mathbf{corr}(X,Y|Z_1,...,Z_m) = \frac{\mathbf{corr}(X,Y|Z_2,...,Z_m) - \mathbf{corr}(X,Z_1|Z_2,...,Z_m)\,\mathbf{corr}(Y,Z_1|Z_2,...,Z_m)}{\sqrt{(1 - \mathbf{corr}(X,Z_1|Z_2,...,Z_m))(1 - \mathbf{corr}(Y,Z_1|Z_2,...,Z_m))}}$$

Для оценки коэффициента корреляции естественно использовать выборочный коэффициент корреляции переменных X и Y на основе наблюдений (X_1,Y_1),....,(X_n,Y_n)

$$\mathbf{Corr}(X,Y) = \frac{\mathbf{Cov}(X,Y)}{S_X S_Y},$$

где $S_X^2 = \overline{X^2} - \overline{X}^2$, $S_Y^2 = \overline{Y^2} - \overline{Y}^2$ — выборочные дисперсии, $\mathbf{Cov}(X,Y)$ — выборочная ковариация, равная $\overline{XY} - \overline{X}\,\overline{Y}$.

Аналогичным образом частный выборочный коэффициент X и Y при условии Z на основе выборок (X_1,Y_1,Z_1),...,(X_n,Y_n,Z_n) равен

$$\mathbf{Corr}(X,Y|Z) = \frac{\mathbf{Corr}(X,Y) - \mathbf{Corr}(Y,Z)\,\mathbf{Corr}(X,Z)}{\sqrt{(1 - \mathbf{Corr}(X,Z))\,(1 - \mathbf{Corr}(Y,Z))}}.$$

При условии набора выборок $Z_{\cdot,1},...,Z_{\cdot,m}$ мы опять же можем определять коэффициент рекуррентно

$$\mathbf{Corr}(X,Y|Z_{\cdot,1},...,Z_{\cdot,m}) = \frac{\mathbf{Corr}(X,Y|Z_{\cdot,2},...,Z_{\cdot,m}) - \mathbf{Corr}(X,Z_{\cdot,1}|Z_{\cdot,2},...,Z_{\cdot,m})\,\mathbf{Corr}(Y,Z_{\cdot,1}|Z_{\cdot,2},...,Z_{\cdot,m})}{\sqrt{(1 - \mathbf{Corr}(X,Z_{\cdot,1}|Z_{\cdot,2},...,Z_{\cdot,m}))(1 - \mathbf{Corr}(Y,Z_{\cdot,1}|Z_{\cdot,2},...,Z_{\cdot,m}))}}$$

или же из соображений линейности, подсчитывая выборочный коэффициент корреляции $X - E_X$, $Y - E_Y$, где

$$E_X = a_{X,1}Z_{\cdot,1} + ... + a_{X,m}Z_{\cdot,m} + a_X(1,...,1),\ E_Y = a_{Y,1}Z_{\cdot,1} + ... + a_{Y,m}Z_{\cdot,m} + a_Y(1,...,1).$$

0.2 Регрессия и корреляция

Пример 2. Напомню, что мы получили F-критерий

$$\frac{D_{0,1}/(m-m_0)}{D_1/(n-m)} = \frac{(D_0-D_1)/(m-m_0)}{D_1/(n-m)} > f_{1-\alpha},\ D_{0,1} = ||\operatorname{proj}_L \vec{y} - \operatorname{proj}_{L_0} \vec{y}||^2,\ D_1 = ||\vec{y} - \operatorname{proj}_L \vec{y}||^2,$$

где f — квантиль распределения Фишера-Снедекора с $m-m_0, n-m$ степенями свободы. Для примера рассмотрим гипотезу $H_0 : Y_i = b+\varepsilon_i$ с $H_0 \cup H_1 : Y_i = ax_i+b+\varepsilon_i$. "Правдоподобность" гипотезы отражает коэффициент

$$R^2 = 1 - \frac{D_1}{D_0},$$

который меняется от 0 до 1 и близость его к 1 означает, что гипотеза крайне маловероятна и данные плохо описываются горизонтальной прямой. При этом

$$D_1 = \sum_{i=1}^n (y_i - \widehat{a}x_i - \widehat{b})^2 = \sum_{i=1}^n (y_i - \overline{y} - \widehat{a}(x_i - \overline{x}))^2 = \sum_{i=1}^n (y_i-\overline{y})^2 + \widehat{a}^2 \sum_{i=1}^n (x_i-\overline{x})^2 - 2\widehat{a}\sum_{i=1}^n (y_i-\overline{y})(x_i-\overline{x}) =$$
$$\sum_{i=1}^n (y_i-\overline{y})^2 - \frac{\left(\sum_{i=1}^n (y_i-\overline{y})(x_i-\overline{x})\right)^2}{\sum_{i=1}^n (x_i-\overline{x})^2},\ D_0 = \sum_{i=1}^n (y_i-\overline{y})^2,$$

откуда

$$R^2 = \frac{\left(\sum_{i=1}^n (y_i-\overline{y})(x_i-\overline{x})\right)^2}{\sum_{i=1}^n (x_i-\overline{x})^2 \sum_{i=1}^n (y_i-\overline{y})^2}.$$

Таким образом, коэффициент R^2 есть квадрат выборочного коэффициента корреляции $Corr(\vec{x}, \vec{y})$.

В более общем случае рассмотрим задачу снижения размерности. Пусть мы знаем, что $Y = X\vec{a} + \vec{\varepsilon}$ и хотим посмотреть на модель $Y = \widetilde{X}\vec{b} + \vec{\varepsilon}$, $\widetilde{X}$ — матрица X без первого столбца. Будем для удобства считать, что $X_{\cdot,m}$ есть вектор $(1, 1..., 1)$. Тогда в первом случае RSS является расстоянием от $\vec{Y}$ до пространства, порожденного $X_{\cdot,1}$,...$X_{\cdot,m}$, а во втором — до пространства, порожденного $X_{\cdot,2}$,...,$X_{\cdot,m}$. Следовательно, разность квадратов расстояний $D_{0,1} = D_0 - D_1$ будет равна квадрату длины проекции $\vec{Y}$ на $L_1 \cap L_0^\perp$ — ортогональное направление к $L_0^\perp$ в L_1.

Сформулируем вышесказанной в терминах прошлого занятия (последовательно ортогонализированных предикторов). Пусть $\vec{Z}_m$ — ортогонализация $X_{\cdot,1}$ относительно $X_{\cdot,2}$,...,$X_{\cdot,m}$ или, другими словами, проекция $X_{\cdot,1}$ на перпендикуляр к пространству L_0, порожденному $X_{\cdot,2}$,...,$X_{\cdot,m}$, в пространстве L_1. Пусть $\vec{Y}_m$ аналогичным образом остаток регрессии $\vec{Y}$ на $X_{\cdot,2}$,...,$X_{\cdot,m}$, то есть проекция $\vec{Y}$ на ортогональное дополнение к L_0 в $\mathbb{R}^n$. Тогда $D_{0,1}$ — квадрат длины проекции $\vec{Y}_m$ на $\vec{Z}_m$, D_0 — квадрат длины проекции $\vec{Y}_m$, иначе говоря

$$\frac{D_{0,1}}{D_0} = \frac{\langle \vec{Y}_m, \vec{Z}_m\rangle^2}{||\vec{Y}_m||^2||\vec{Z}_m||^2} = \operatorname{Corr}(\vec{Y}_m, \vec{Z}_m)^2,$$

где мы воспользовались тем, что $\overline{\vec{Y}_m} = 0$, $\overline{\vec{Z}_m} = 0$, поскольку оба вектора ортогональны $(1, ..., 1)$. Итак, аналог R^2 в этом случае будет представлять собой квадрат частного выборочного коэффициента корреляции $X_{\cdot,1}$ и $\vec{Y}$ при условии $X_{\cdot,2}$,...,$X_{\cdot,m}$.

Пример 3. Хукер (1907 год) в нескольких графствах Англии рассматривал урожайность x и сумму температур z выше 5.5 градусов за весну. Оказалось, что $\rho_{x,y} = -0.4$. Правда ли высокая температура весной отрицательно влияет на урожайность? Картина меняется, если рассмотреть дополнительный параметр — весеннее количество осадков z. Коэффициенты корреляции $\rho_{x,z} = 0.8$, $\rho_{y,z} = -0.56$, откуда

$$\mathbf{Corr}(x, y|z) = \frac{-0.4 + 0.8 \cdot 0.56}{\sqrt{1-0.8^2}\sqrt{1-0.56^2}} \approx 0.09.$$

Таким образом, после исключения уровня осадков корреляция x и y стала небольшой и положительной.