

2 Linear Regression (theory)

1. Let $Z = (X', W')'$. Then,

$$E^*[Y|X, W] = E^*[Y|Z] = E[Y] + \sigma(Y, Z)V(Z)^{-1}(Z - E[Z])$$

Because W and X are independent, $V(Z) = \begin{pmatrix} V(X) & 0 \\ 0 & V(W) \end{pmatrix}$. So,

$$\begin{aligned} E^*[Y|X, W] &= E[Y] + (\sigma(Y, X) \quad \sigma(Y, W)) \begin{pmatrix} V(X)^{-1} & 0 \\ 0 & V(W)^{-1} \end{pmatrix} \begin{pmatrix} X - E[X] \\ W - E[W] \end{pmatrix} \\ &= E[Y] + (\sigma(Y, X) \quad \sigma(Y, W)) \begin{pmatrix} V(X)^{-1}(X - E[X]) \\ V(W)^{-1}(W - E[W]) \end{pmatrix} \\ &= E[Y] + \sigma(Y, X)V(X)^{-1}(X - E[X]) + \sigma(Y, W)V(W)^{-1}(W - E[W]) \\ &= E[Y] + E^*[Y|X] - E[Y] + E^*[Y|W] - E[Y] \\ &= E^*[Y|X] + E^*[Y|W] - E[Y] \end{aligned}$$

2. Using the formula for the best linear predictor β_0

$$|\beta_0| = \frac{|\sigma(X, Y)|}{V(X)} \leq \frac{\sqrt{V(X)V(Y)}}{V(X)} = \frac{\sqrt{V(Y)}}{\sqrt{V(X)}} = 1$$

In the example of father's height and child's height, the interpretation is that taller fathers may have taller children, but these effects cannot be amplified between generations. The expectation of the child's height will be the average height plus some fraction of the deviation of the father's height from the average.

3. From the formula of variance and applying the Law of Iterated Expectations,

$$\begin{aligned} V(Y) &= E[Y^2] - E[Y]^2 \\ &= E[E[Y^2|X]] - E[E[Y|X]]^2 \\ &= E[V(Y|X) + E[Y|X]^2] - E[E[Y|X]]^2 \\ &= E[V(Y|X)] + E[E[Y|X]^2] - [E[E[Y|X]]]^2 = E[V(Y|X)] + V(E[Y|X]) \end{aligned}$$

4. (a) Using the formula for the best linear predictor, $E^*[m(Z)|X] = \delta_0 + \xi_0 X$, and the law of iterated expectations such that $E[m(Z)] = E[E[X|Z]] = E[X]$,

$$\begin{aligned} \xi_0 &= \sigma(m(Z), X)V(X)^{-1} \\ &= E \left[\left(m(Z) - E[m(Z)] \right) \left(X - E[X] \right) \right] \left(E[V(X|Z)] + V(E[X|Z]) \right)^{-1} \\ &= E \left[\left(E[X|Z] - E[X] \right) \left(X - E[X] \right) \right] \left(E[V(X|Z)] + V(E[X|Z]) \right)^{-1} \\ &= E \left[\left(E[X|Z] - E[E[X|Z]] \right) \left(X - E[E[X|Z]] \right) \right] \left(E[V(X|Z)] + V(E[X|Z]) \right)^{-1} \\ &= E \left[\left(E[X|Z] - E[E[X|Z]] \right)^2 \right] \left(E[V(X|Z)] + V(E[X|Z]) \right)^{-1} \\ &= V(E[X|Z]) \left(E[V(X|Z)] + V(E[X|Z]) \right)^{-1} \end{aligned}$$

By the best linear predictor, $\delta_0 = E[m(Z)] - \xi_0 E[X] = (1 - \xi_0)E[X]$.

- (b) ξ_0 is the fraction of total variance of the parents' income that comes from variation of average income between individual neighborhoods, with the rest of the variance due to variation of income within individual neighborhoods. That is, the higher ξ_0 is, the more stratified residential areas are by income.

- (c) First note that $\rho = \frac{\sigma(A, X)}{\sigma_A \sigma_X}$, and that the best linear predictor of A on X is given by

$$\begin{aligned} E^*[A|X] &= \mu_A + \sigma(A, X)(\sigma_A^2)^{-1}(X - \mu_x) \\ &= \mu_A + \frac{\rho\sigma_A}{\sigma_X}(X - \mu_x) \end{aligned}$$

So, taking the linear prediction of Y with respect to X and substituting the results from above:

$$\begin{aligned} E^*[Y|X, m(Z), A] &= \alpha_0 + \beta_0 X + \gamma_0 m(Z) + A \\ E^*[Y|X] &= \alpha_0 + \beta_0 X + \gamma_0 E^*[m(Z)|X] + E^*[A|X] \\ E^*[Y|X] &= \alpha_0 + \beta_0 X + \gamma_0((1 - \xi_0)\mu_X + \xi_0 X) + \mu_A + \frac{\rho\sigma_A}{\sigma_X}(X - \mu_x) \\ E^*[Y|X] &= \alpha_0 + \gamma_0(1 - \xi_0)\mu_X + \left(\mu_A - \frac{\rho\sigma_A}{\sigma_X}\mu_x\right) + \left(\beta_0 + \gamma_0\xi_0 + \rho\frac{\sigma_A}{\sigma_X}\right)X \end{aligned}$$

- (d) If there is more income stratification in New York than San Francisco, $\xi_{NYC} > \xi_{SF}$. So, the intercept for New York will be lower than that of San Francisco, but the coefficient on X will be higher in New York.

3 Hansen textbook problems

2.16

First, the marginal distributions of x and y are given by

$$\begin{aligned} f(x) &= \int_0^1 \frac{3}{2}(x^2 + y^2)dy \\ &= \frac{3}{2}(x^2 y + \frac{1}{3}y^3) \Big|_0^1 = \frac{3}{2}x^2 + \frac{1}{2} \\ f(y) &= \frac{3}{2}y^2 + \frac{1}{2} \end{aligned}$$

So, $E[x] = E[y] = \frac{5}{8}$, and $V(x) = V(y) = \frac{7}{15} - \frac{25}{64} = \frac{73}{960}$. The covariance of x and y is given by

$$\begin{aligned} \sigma(x, y) &= E[xy] - E[x]E[y] = \int_0^1 \int_0^1 xy \frac{3}{2}(x^2 + y^2)dydx - \left(\frac{5}{8}\right)^2 \\ &= \int_0^1 \int_0^1 \frac{3}{2}(x^3 y + xy^3)dydx - \frac{25}{64} \\ &= \int_0^1 \left[\frac{3}{4}x^3 y^2 + \frac{3}{8}xy^4 \right]_0^1 dx - \frac{25}{64} \\ &= \int_0^1 \left(\frac{3}{4}x^3 + \frac{3}{8}x \right) dx - \frac{25}{64} \\ &= \frac{3}{16} + \frac{3}{16} - \frac{25}{64} = -\frac{1}{64} \end{aligned}$$

The best linear predictor of y will be given by $E[y|x] = \alpha + \beta x$, with

$$\beta = \sigma(x, y)V(x)^{-1} = -\frac{1}{64} \frac{960}{73} = -\frac{15}{73}$$

And $\alpha = E[y] - E[x]\beta = (1 + \frac{15}{73})\frac{5}{8}$.

The conditional mean of y given x is

$$\begin{aligned} E[y|x] &= \frac{\int_0^1 \frac{3}{2}(x^2 + y^2)y dy}{\frac{3}{2}x^2 + \frac{1}{2}} \\ &= \frac{\int_0^1 \frac{3}{2}(x^2 y + y^3) dy}{\frac{3}{2}x^2 + \frac{1}{2}} = \frac{\frac{3}{4}x^2 + \frac{3}{8}}{\frac{3}{2}x^2 + \frac{1}{2}} \end{aligned}$$

Which is clearly not the best linear predictor.

2.18

(a) Using $x_3 = \alpha_1 + \alpha_2 x_2$,

$$\begin{aligned} Q_{xx} = E[(xx')] &= E\begin{bmatrix} 1 & x_2 & x_3 \\ x_2 & x_2^2 & x_2 x_3 \\ x_3 & x_2 x_3 & x_3^2 \end{bmatrix} \\ &= E\begin{bmatrix} 1 & x_2 & (\alpha_1 + \alpha_2 x_2) \\ x_2 & x_2^2 & x_2(\alpha_1 + \alpha_2 x_2) \\ \alpha_1 + \alpha_2 x_2 & x_2(\alpha_1 + \alpha_2 x_2) & (\alpha_1 + \alpha_2 x_2)^2 \end{bmatrix} \\ &= E\begin{bmatrix} 1 & x_2 & (\alpha_1 + \alpha_2 x_2) \\ x_2 & x_2^2 & x_2(\alpha_1 + \alpha_2 x_2) \\ \alpha_1 + \alpha_2 x_2 & \alpha_1 x_2 + \alpha_2 x_2^2 & \alpha_1(\alpha_1 + \alpha_2 x_2) + \alpha_2(x_2(\alpha_1 + \alpha_2 x_2)) \end{bmatrix} \end{aligned}$$

Where the third row is a linear combination of the first two (α_1 of the first and α_2 of the second), so this is not invertible.

(b) Taking the linear transformation Ax such that $A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$, I can apply ordinary least squares

to find a γ such that $\gamma = \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix}$ are the coefficients on 1 and x_2 for the best linear predictor of y given 1 and x_2 . Then, the best linear predictor of y given x is $(\gamma_1 \ \gamma_2 \ \gamma_3)A = (\gamma_1 \ \gamma_2 \ 0)$. That is, the best linear predictor of y given x is also the best linear predictor of y given $(1, x_2)$.

2.19

Rewriting $d(\beta) = E(m(x) - x'\beta)^2 = E(m(x)^2 - 2\beta'xm(x) + \beta'x'x\beta)$, the first-order condition is

$$\begin{aligned} E(xx')\beta &= E(xm(x)) \\ \beta &= E(xx')^{-1}E(xm(x)) \\ \beta &= E(xx')^{-1}E(xm(x)) \\ \beta &= E(xx')^{-1}E(xE[y|x]) \\ \beta &= E(xx')^{-1}E(E[xy|x]) \\ \beta &= E(xx')^{-1}E(xy) \end{aligned}$$

3.7

Show that if $X = [X_1 X_2]$, then $PX_1 = X_1$ and $MX_1 = 0$: Denote the dimensions of X_1 and X_2 as $N \times K_1$ and $N \times K_2$, respectively. Then, the dimensionality of P is $N \times N$, and $PX = P[X_1 X_2] = [PX_1 PX_2] = [X_1 X_2]$, so $PX_1 = X_1$. Then, $MX_1 = (I_N - P)(X_1) = X_1 - PX_1 = X_1 - X_1 = 0$.

3.8

Show that $MM = M$:

$$\begin{aligned} MM &= (I - X(X'X)^{-1}X')(I - X(X'X)^{-1}X') \\ &= I - 2X(X'X)^{-1}X' + X(X'X)^{-1}X'X(X'X)^{-1}X' \\ &= I - 2X(X'X)^{-1}X' + X(X'X)^{-1}X' = I - P = M \end{aligned}$$

3.9

Show that $\text{tr}(M) = n - k$:

$$\begin{aligned} \text{tr}(M) &= \text{tr}(I - P) = \text{tr}(I) - \text{tr}(P) \\ &= n - \text{tr}(X(X'X)^{-1}X') \\ &= n - \text{tr}(X'X(X'X)^{-1}) \\ &= n - \text{tr}(I_k) = n - k \end{aligned}$$

3.10

Show that $P = P_1 + P_2$ given $X_1'X_2 = 0$: Note that $X_1'X_2 = 0 \implies X_2'X_1 = 0$.

$$\begin{aligned} P &= (X_1 \ X_2) \begin{pmatrix} X_1'X_1 & 0 \\ 0 & X_2'X_2 \end{pmatrix}^{-1} \begin{pmatrix} X_1' \\ X_2' \end{pmatrix} \\ &= (X_1 \ X_2) \begin{pmatrix} (X_1'X_1)^{-1} & 0 \\ 0 & (X_2'X_2)^{-1} \end{pmatrix} \begin{pmatrix} X_1' \\ X_2' \end{pmatrix} \\ &= (X_1(X_1'X_1)^{-1} \ X_2(X_2'X_2)^{-1}) \begin{pmatrix} X_1' \\ X_2' \end{pmatrix} \\ &= X_1(X_1'X_1)^{-1}X_1' + X_2(X_2'X_2)^{-1}X_2' = P_1 + P_2 \end{aligned}$$

5.2

1. Show that \bar{y}^* is unbiased:

$$E[\bar{y}^*] = E\left[\frac{1}{n} \sum_{i=1}^n w_i y_i\right] = \frac{1}{n} \sum_{i=1}^n w_i E[y_i] = \mu$$

2. Calculate $V(\bar{y}^*)$:

$$V(\bar{y}^*) = V\left(\frac{1}{n} \sum_{i=1}^n w_i y_i\right) = \frac{1}{n^2} \sum_{i=1}^n w_i^2 V(y_i) = \frac{\sum_{i=1}^n w_i^2}{n^2} V(y_i)$$

3. Using Chebyshev's Inequality and some $\delta > 0$

$$Pr(|\bar{y}_n^* - \mu| > \delta) \leq \frac{V(\bar{y}_n^*)}{\delta^2} = \frac{\sum_{i=1}^n w_i^2 / n^2}{\delta^2}$$

If the numerator goes to 0, then $\bar{y}_n^* \xrightarrow{P} \mu$.

4. Show that a sufficient condition for the condition in part 3 is $\max_{i \leq n} w_i = o(n)$: The condition from part 3 was that the variance of $\bar{y}_n^* \xrightarrow{P} 0$. Let $w = \max_{i \leq n} w_i = o(n)$. Noting that $w \leq 1 \implies w^2 \leq w$,

$$\begin{aligned} \frac{\sum_{i=1}^n w_i^2}{n^2} V(y_i) &\leq \frac{\sum_{i=1}^n w^2}{n^2} V(y_i) \\ &= \frac{w^2}{n} V(y_i) < \frac{w}{n} V(y_i) \end{aligned}$$

Because $w = o(n)$, this term goes to 0, so the variance goes to 0 as well.

5.3

Applying Chebyshev's Inequality,

$$Pr(|Z| > \delta) \leq \frac{1}{\delta^2} = 0.05 \implies \delta = \frac{1}{\sqrt{0.05}} \approx 4.47$$

If Z is standard normal, then $\delta = 1.96$. So, Chebyshev's inequality is very, very conservative relative to the actual value.