Ahmad Zareei
Econometrics 240A
Homework 3 - Part II

Problem 1. Linear regression: theory:

[a]. Assume that (i) $\mathbb{E}[Y^2] < \infty$, (ii) $\mathbb{E}[||X||^2] < \infty$ and (iii) $\mathbb{E}[(\alpha' X)^2] > 0$ for any non-zero $\alpha \in \mathbb{R}^K$. Let $X'b$ be a linear predictor of Y given X. Let $U = Y - X'\beta_0$ and show that

$$\mathbb{E}[(Y - X'b)^2] = \mathbb{E}[U]^2 + 2(\beta_0 - b)'\mathbb{E}[XU] + (\beta_0 - b)'\mathbb{E}[XX'](\beta_0 - b) \tag{1}$$

Solution: We have

$$\begin{aligned}
\mathbb{E}[(Y - X'b)^2] &= \mathbb{E}[(Y - X'\beta + X'\beta - X'b)^2] \\
&= \mathbb{E}[(U + X'(\beta - X'b))^2] \\
&= \mathbb{E}[U^2 + 2(\beta_0 - b)'XU + (beta_0 - b)'XX'(\beta_0 - b)] \\
&= \mathbb{E}[U^2] + 2(\beta_0 - b)'\mathbb{E}[XU] + (beta_0 - b)'\mathbb{E}[XX'](\beta_0 - b)
\end{aligned} \tag{2}$$

where it proves the statement.

[b]. Show that if $\mathbb{E}[XU] = 0$ (you may assume that X includes a constant), then

$$\mathbb{E}[(Y - X'b)^2] \geq E[U^2] \tag{3}$$

with strict inequality unless $b = \beta_0$.

Solution. Using statement we found in part [a], we set $\mathbb{E}[XU] = 0$, we obtain

$$\begin{aligned}
\mathbb{E}[(Y - X'b)^2] &= \mathbb{E}[U^2] + 2(\beta_0 - b)'\mathbb{E}[XU] + (beta_0 - b)'\mathbb{E}[XX'](\beta_0 - b) \\
&= \mathbb{E}[U^2] + (beta_0 - b)'\mathbb{E}[XX'](\beta_0 - b) \qquad \text{since } \mathbb{E}[XU] = 0 \\
&\geq \mathbb{E}[U^2] \qquad \mathbb{E}[XX'] > 0, \text{ proof bellow and if } b = \beta_0 \text{ the equality holds}
\end{aligned} \tag{4}$$

where in the last line of the above proof, we used the fact that $\mathbb{E}[XX']?0$, since $\mathbb{E}[(\alpha' X)^2] > 0, \quad \forall \alpha \in \mathbb{R}^K$, we have

$$\forall \alpha \in \mathbb{R}^K \qquad \mathbb{E}[(\alpha' X)^2] > 0 \leftrightarrow \mathbb{E}[\alpha' XX'\alpha] > 0 \leftrightarrow \alpha'\mathbb{E}[XX']\alpha > 0 \leftrightarrow \mathbb{E}[XX'] > 0 \tag{5}$$

[c]. (Pythagorean Rule) Show that

$$\mathbb{V}(Y) = \mathbb{V}(Y - \mathbb{E}^*[Y|X]) + \mathbb{V}(\mathbb{E}^*[Y|X]) \tag{6}$$

Solution. We define $U = Y - \mathbb{E}^*[Y|X]$, where we have $\mathbb{E}[XU] = 0$, and $\mathbb{E}[U] = 0$ (X contains 1 in the vector form). Therefore the covariance of $\mathbb{E}^*[Y|X]$, and U becomes zero, i.e. $\mathbb{C}(\mathbb{E}^*[Y|X], U) = 0$. Then

$$\begin{aligned}\mathbb{V}(Y) &= \mathbb{V}(U + \mathbb{E}^*[Y|X]) = \mathbb{V}(U) + \mathbb{V}(\mathbb{E}^*[Y|X]) + 2\mathbb{C}(\mathbb{E}^*[Y|X], U) \\ &= \mathbb{V}(U) + \mathbb{V}(\mathbb{E}^*[Y|X]) \\ &= \mathbb{V}(Y - \mathbb{E}^*[Y|X]) + \mathbb{V}(\mathbb{E}^*[Y|X]) \end{aligned} \tag{7}$$

[d]. Let $X_1, ..., X_K$ be a set of regressors with the property that $\mathbb{C}(X_k, X_l) = 0$, for all $k \neq l$. Show that

$$\mathbb{E}^*[Y|X_1, ..., X_K] = \sum_{k=1}^{K} \mathbb{E}^*[Y|X_k] - (K-1)\mathbb{E}[Y] \tag{8}$$

Solution. We know that

$$\mathbb{E}^*[Y|X_k] = \alpha_k + \beta_k X_k, \qquad \beta_k = \frac{\mathbb{C}(Y, X_k)}{\mathbb{V}(X_k}, \quad \alpha_k = \mathbb{E}[Y] - \beta_k \mathbb{E}[X_k] \tag{9}$$

We know proove that $\mathbb{E}[UX_l] = 0$, we have

$$\begin{aligned}\mathbb{E}[UX_l] &= \mathbb{E}\left[\left(Y - \sum_{k=1}^{K} \mathbb{E}^*[Y|X_k] + (K-1)\mathbb{E}[Y]\right) X_l\right] \\ &= \mathbb{E}\left[\left(Y - \mathbb{E}[Y] - \sum_{k=1}^{K} \left(\mathbb{E}^*[Y|X_k] - \mathbb{E}[Y]\right)\right) X_l\right] \\ &= \mathbb{E}\left[\left(Y - \mathbb{E}[Y]\right) X_l\right] - \sum_{k=1}^{K} \mathbb{E}\left[\left(\mathbb{E}^*[Y|X_k] - \mathbb{E}[Y]\right) X_l\right] \\ &= \mathbb{E}\left[\left(Y - \mathbb{E}[Y]\right) \left(X_l - \mathbb{E}[X_l]\right)\right] - \sum_{k=1}^{K} \mathbb{E}\left[\left(\alpha_k + \beta_k X_k - \mathbb{E}[Y]\right) X_l\right] \\ &= \mathbb{C}(Y, X_l) - \sum_{k=1}^{K} \mathbb{E}\left[\left(\mathbb{E}[Y] - \beta_k \mathbb{E}[X_k] + \beta_k X_k - \mathbb{E}[Y]\right) X_l\right] \\ &= \mathbb{C}(Y, X_l) - \sum_{k=1}^{K} \mathbb{E}\left[\beta_k \left(X_k - \mathbb{E}[X_k]\right) X_l\right] \\ &= \mathbb{C}(Y, X_l) - \sum_{k=1}^{K} \beta_k \mathbb{E}\left[\left(X_k - \mathbb{E}[X_k]\right) \left(X_l - \mathbb{E}[X_l]\right)\right] \\ &= \mathbb{C}(Y, X_l) - \sum_{k=1}^{K} \beta_k \mathbb{C}(X_k, X_l) \\ &= \mathbb{C}(Y, X_l) - \beta_l \mathbb{C}(X_l, X_l) \\ &= \mathbb{C}(Y, X_l) - \beta_l \mathbb{V}(X_l) \\ &= \mathbb{C}(Y, X_l) - \mathbb{C}(Y, X_l) = 0 \end{aligned} \tag{10}$$

since this is true for every l, this would mean that $\mathbb{E}[XU] = 0$. Then the results follows using the result of the previous parts.

[e] Under the same conditions in part [c] above show that

$$\mathbb{E}^*[Y|X_1, ..., X_K] = \mathbb{E}[Y] + \sum_{k=1}^{K} \frac{\mathbb{C}(Y, X_k)}{\mathbb{V}(X_k)}(X_k - \mathbb{E}[X_k]) \tag{11}$$

and hence that the proportion of variance 'explained' equals

$$1 - \frac{\mathbb{V}(U)}{\mathbb{V}(Y)} = \sum_{k=1}^{K} \rho_k^2 \tag{12}$$

where $\rho_k = \mathbb{C}(Y, X_k)/\left(\mathbb{V}(X_k)^{1/2}\mathbb{V}(Y)^{1/2}\right)$.

Solution. Using the results of part [d], we have

$$\begin{aligned}
\mathbb{E}^*[Y|X_1, ..., X_K] &= \sum_{k=1}^{K} \mathbb{E}^*[Y|X_k] - (K-1)\mathbb{E}[Y] \\
&= \mathbb{E}[Y] + \sum_{k=1}^{K} (\mathbb{E}^*[Y|X_k] - \mathbb{E}[Y]) \\
&= \mathbb{E}[Y] + \sum_{k=1}^{K} (\alpha_k + \beta_k X_k - \mathbb{E}[Y]) \\
&= \mathbb{E}[Y] + \sum_{k=1}^{K} (\mathbb{E}[Y] - \beta_k \mathbb{E}[X_k] + \beta_k X_k - \mathbb{E}[Y]) \\
&= \mathbb{E}[Y] + \sum_{k=1}^{K} \beta_k (X_k - \mathbb{E}[X_k]) \\
&= \mathbb{E}[Y] + \sum_{k=1}^{K} \frac{\mathbb{C}(Y, X_k)}{\mathbb{V}(X_k)} (X_k - \mathbb{E}[X_k])
\end{aligned} \tag{13}$$

Now for the second part, we first need to find the variance of the above term, we have

$$\begin{aligned}
\mathbb{V}(\mathbb{E}^*[Y|X_1, ..., X_K]) &= \mathbb{V}\left(\mathbb{E}[Y] + \sum_{k=1}^{K} \frac{\mathbb{C}(Y, X_k)}{\mathbb{V}(X_k)} (X_k - \mathbb{E}[X_k])\right) \\
&= \mathbb{V}\left(\sum_{k=1}^{K} \frac{\mathbb{C}(Y, X_k)}{\mathbb{V}(X_k)} X_k\right) \\
&= \sum_{k=1}^{K} \left(\frac{\mathbb{C}(Y, X_k)}{\mathbb{V}(X_k)}\right)^2 \mathbb{V}(X_k) + \sum_{k \neq j} \frac{\mathbb{C}(Y, X_k)}{\mathbb{V}(X_k)} \frac{\mathbb{C}(Y, X_l)}{\mathbb{V}(X_l)} \mathbb{C}(X_k, X_l) \\
&= \sum_{k=1}^{K} \frac{\mathbb{C}(Y, X_k)^2}{\mathbb{V}(X_k)} \qquad \text{since } \mathbb{C}(X_k, X_l) = 0
\end{aligned} \tag{14}$$

Therefore, we have

$$\begin{aligned}
1 - \frac{\mathbb{V}(U)}{\mathbb{V}(Y)} &= \frac{\mathbb{V}(Y) - \mathbb{V}(U)}{\mathbb{V}(Y)} \\
&= \frac{\mathbb{V}(\mathbb{E}^*[Y|X_1, ..., X_K]}{\mathbb{V}(Y)} \qquad \text{part c} \\
&= \sum_{k=1}^{K} \frac{\mathbb{C}(Y, X_k)^2}{\mathbb{V}(X_k)\mathbb{V}(Y)} \\
&= \sum_{k=1}^{K} \rho_k^2
\end{aligned} \tag{15}$$

Problem 2. Linear Regression, Application #1

Let $m(Z) = \mathbb{E}[X|Z]$ and consider the linear regression as

$$\mathbb{E}^*[Y|X, m(Z), A] = \alpha_0 + \beta_0 X + \gamma_0 m(Z) + A \tag{16}$$

[a]. Show that

$$\mathbb{E}^*[m(Z)|X] = \delta_0 + \zeta_0 X \tag{17}$$

with

$$\delta_0 = (1 - \zeta_0)\mathbb{E}[X] \tag{18}$$

$$\zeta_0 = \frac{\mathbb{V}(\mathbb{E}[X|Z])}{\mathbb{E}[\mathbb{V}(X|Z)] + \mathbb{V}(\mathbb{E}[X|Z])} \tag{19}$$

Solution. We have

$$\delta_0 = \mathbb{E}[m(Z)] - \zeta_0 \mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Z]] - \zeta_0 \mathbb{E}[X] = (1 - \zeta_0)\mathbb{E}[X] \tag{20}$$

and also

$$\begin{aligned}
\zeta_0 &= \frac{\mathbb{C}(m(Z), X)}{\mathbb{V}(X)} \\
&= \frac{\mathbb{C}(m(Z), X)}{\mathbb{E}[\mathbb{V}(X|Z)] + \mathbb{V}(\mathbb{E}[X|Z])} \qquad \text{since } \mathbb{V}(X) = \mathbb{E}[\mathbb{V}(X|Z)] + \mathbb{V}(\mathbb{E}[X|Z]) \\
&= \frac{\mathbb{C}(\mathbb{E}[X|Z], X)}{\mathbb{E}[\mathbb{V}(X|Z)] + \mathbb{V}(\mathbb{E}[X|Z])} \\
&= \frac{\mathbb{C}(\mathbb{E}[X|Z], \mathbb{E}[X|Z] + e)}{\mathbb{E}[\mathbb{V}(X|Z)] + \mathbb{V}(\mathbb{E}[X|Z])} \qquad \text{where } X = \mathbb{E}[X|Z] + e, \quad \mathbb{E}[e|Z] = 0, \quad \mathbb{E}[e] = 0 \\
&= \frac{\mathbb{V}(\mathbb{E}[X|Z])}{\mathbb{E}[\mathbb{V}(X|Z)] + \mathbb{V}(\mathbb{E}[X|Z])}
\end{aligned} \tag{21}$$

where we used the fact that

$$\mathbb{C}(\mathbb{E}[X|Z], e) = \mathbb{E}[\mathbb{E}[X|Z]e] - \mathbb{E}[\mathbb{E}[X|Z]]\mathbb{E}[e] = \mathbb{E}[\mathbb{E}[X|Z]e] = \mathbb{E}\left[\mathbb{E}[\mathbb{E}[X|Z]\mathbb{E}[e|Z]]\right] = 0 \tag{22}$$

[b]. Assume the population under consideration is working age adults who grew up in the San Francisco Bay Area. Let Y denote a adult log income, let X denote the log income of ones parents as a child and let Z be a vector of dummy variables denoting an individuals neighborhood of residence as a child. Provide an interpretation of ζ_0 as a measure of residential stratification by income.

Solution. Here ζ_0 depends on two components that are $\mathbb{E}[\mathbb{V}(X|Z)]$ and $\mathbb{V}(\mathbb{E}[X|Z])$, where X denotes the log income of ones paretns and Z is a vector of dummy variables denoting an individual neighborhood of resdience. The first term, $\mathbb{E}[\mathbb{V}(X|Z)]$ tells us the mean dispersion of parents' earnings in each neighborhood and teh second term, $\mathbb{V}(\mathbb{E}[X|Z])$, tells us how unequal the distribution of parents' earnings is across different neighborhoods. Here $\mathbb{V}(\mathbb{E}[X|Z])$ is a measure of stratification of income among residential areas, the higher values shows higher inequality within a city. Therefore ζ_0 captures how much of the overall inequality in terms of parental income is due to residential stratification.

[c]. Establish the notation $\rho = \text{corr}(A, X)$, $\mu_A = \mathbb{E}[A]$, $\mu_X = \mathbb{E}[X]$, $\sigma_A^2 = \mathbb{V}(A)$, and $\sigma_X^2 = \mathbb{V}(X)$. Show that

$$\mathbb{E}^*[Y|X] = \alpha_0 + \gamma_0(1-\zeta_0)\mu_X + \left(\mu_A - \rho\frac{\sigma_A}{\sigma_X}\right) + \left[\beta_0 + \gamma_0\zeta_0 + \rho\frac{\sigma_A}{\sigma_X}\right]. \tag{23}$$

Solution. We have

$$Y = \alpha_0 + \beta_0 X + \gamma_0 m(Z) + A + e \tag{24}$$

where e denotes the error and $\mathbb{E}[e] = 0$ and $\mathbb{C}(X, e) = \mathbb{C}(m(Z), e) = \mathbb{C}(A, e) = 0$. Therefore we obtain

$$\begin{aligned}
\mathbb{E}^*[Y|X] &= \alpha_0 + \beta_0 X + \gamma_0 \mathbb{E}[m(Z)|X] + \mathbb{E}^*[A|X] \\
&= \alpha_0 + \beta_0 X + \gamma_0 \left(\delta_0 + \zeta_0 X\right) + \mathbb{E}^*[A|X] \\
&= \alpha_0 + \beta_0 X + \gamma_0 \left((1-\zeta_0)\mathbb{E}[X] + \zeta_0 X\right) + \mu_A + \frac{\mathbb{C}(X, A)}{\sigma_X^2}(X - \mu_X) \\
&= \alpha_0 + \beta_0 X + \gamma_0(1-\zeta_0)\mu_X + \gamma_0\zeta_0 X + \mu_A + \frac{\rho\sigma_A}{\sigma_X}(X - \mu_X) \\
&= \alpha_0 + \gamma_0(1-\zeta_0)\mu_X + \mu_A + \frac{\rho\sigma_A}{\sigma_X}(X - \mu_X) + \gamma_0\zeta_0 X + \beta_0 X \\
&= \alpha_0 + \gamma_0(1-\zeta_0)\mu_X + \mu_A - \frac{\rho\sigma_A}{\sigma_X}\mu_X + \left(\gamma_0\zeta_0 + \beta_0 + \frac{\rho\sigma_A}{\sigma_X}\right) X
\end{aligned} \tag{25}$$

where the final line is the statement we wanted to proove. Here we used the fact that $\mathbb{E}^*[A|X] = \mu_A + \mathbb{C}(X, A)(X - \mu_X)/\sigma_X^2$.

[d]. Your research assistant computes an estimate of $E^*[Y|X]$ using random sample from San Francisco. She computes a separate estimate using a random sample from New York City. Assume that there is more residential stratification by income in New York than in San Francisco. How would you expect the intercept and slope coefficients to differ across the two regression fits?

Solution. From part [b], we found that the effect of residential stratification is in ζ_0. Using the results in part [c], we can see that the effect of ζ_0 on the intercept is negative, while its effect on the slope is γ_0. Here γ_0 is most likely to be positive, since it represents how expected log-earnings are affected by the average log-earning of the residential neighborhood. So in totoal, higher income stratification, means lower intercept, and higher slope. The intuition is that, poor parents not only have a direct disadvantage, it also has an extra disadvantage by living in poor neighborhoods and hence les opportunity, which means the lower intercept. As parents income increases, not only one gets the direct benefits, he/she gets the indirect benefit of living in a better neighborhood and hence more opportunities, which means positive slope effect.

Problem 2. Linear Regression, Application #2

Solution. In python notebook attached.