Ec240a, Fall 2017

Professor Bryan Graham

Problem Set 3

Due: November 20th, 2017

Problem sets are due at 5PM in the GSIs mailbox. You may work in groups, but each student should turn in their own write-up (including a printout of a narrated/commented and executed Jupyter Notebook if applicable). Please also e-mail a copy of any Jupyter Notebook to the GSI (if applicable).

1 Linear regression: theory

[a] Assume that (i) $\mathbb{E}\left[Y^2\right] < \infty$, (ii) $\mathbb{E}\left[\|X\|^2\right] < \infty$ and (iii) $\mathbb{E}\left[(\alpha' X)^2\right] > 0$ 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

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

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

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

with strict inequality unless $b = \beta_0$.

[c] (Pythagorean Rule) Show that

$$\mathbb{V}\left(Y\right) = \mathbb{V}\left(Y - \mathbb{E}^*\left[Y \middle| X\right]\right) + \mathbb{V}\left(\mathbb{E}^*\left[Y \middle| X\right]\right).$$

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

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

Hint: First show that

$$\mathbb{E}^*\left[\mathbb{E}^*\left[Y \middle| X_k\right] \middle| X_l\right] = \mathbb{E}\left[Y\right]$$

for every $k \neq l$. Second verify the orthogonality conditions

$$\mathbb{E}\left[UX_l\right] = 0$$

for $U = \left(Y - \sum_{k=1}^{K} \mathbb{E}^*\left[Y \middle| X_k\right] + \left(K - 1\right)\mathbb{E}\left[Y\right]\right)$ and $l = 1, \ldots, K$.

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

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

and hence that the proportion of variance 'explained' equals

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

for $\rho_k = \frac{\mathbb{C}(Y,X_k)}{\mathbb{V}(X_k)^{1/2}\mathbb{V}(Y)^{1/2}}$.

2 Linear regression: application #1

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

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

[a] Show that

$$\mathbb{E}^*[m(Z)|X] = \delta_0 + \xi_0 X$$

with

$$\begin{aligned} \delta_0 &= (1-\xi_0)\,\mathbb{E}[X] \\ \xi_0 &= \frac{\mathbb{V}(\mathbb{E}[X|Z])}{\mathbb{E}[\mathbb{V}(X|Z)] + \mathbb{V}(\mathbb{E}[X|Z])}. \end{aligned}$$

[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 one's parents as a child and let Z be a vector of dummy variables denoting an individual's neighborhood of residence as a child. Provide an interpretation of ξ_0 as a measure of residential stratification by income.

[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

$$\mathbb{E}^*[Y|X] = \alpha_0 + \gamma_0(1-\xi_0)\mu_X + \left(\mu_A - \rho\frac{\sigma_A}{\sigma_X}\right) + \left\{\beta_0 + \gamma_0\xi_0 + \rho\frac{\sigma_A}{\sigma_X}\right\} X.$$

[d] Your research assistant computes an estimate of $\mathbb{E}^*[Y|X]$ 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?

3 Linear regression: application #2

The file `brazil_pnad96_ps4.out` contains 65,801 comma delimited records drawn from the 1996 round of the *Brazilian Pesquisas Nacional por Amostra de Domicilos* (PNAD96). The population corresponds to employed males between the ages of 20 and 60. Respondents with incomplete data are dropped from the sample. Each record contains `MONTHLY_EARNINGS, YRSSCH, AgeInDays, Dad_NoSchool_c, Dad_1stPrim_c, Dad_2ndPrim_c, Dad_Sec_c, Dad_DK_c, Mom_NoSchool_c, Mom_1stPrim_c, Mom_2ndPrim_c, Mom_Sec_c, Mom_DK_c` and `ParentsSchooling`. The first three variables equal monthly earnings, years of completed schooling and age in years (but measured to the precision of a day). The next 5 variables are dummies

for father's level of education (no school, first primary cycle completed, second primary cycle completed, secondary or more and 'don't know'). The next 5 variables are the corresponding dummies for mother's level of education. The final variable takes on 25 values corresponding to each possible combination of parent's schooling.

For an analysis of the relationship between schooling and earnings using a closely related dataset you might read the 1991 paper "Declining inequality in schooling in Brazil and its effect on inequality in earnings," by David Lam in the *Journal of Development Economics* 37 (1-2): 199 - 225. This is available of ScienceDirect.

[a] Compute the least squares fit of ln(`MONTHLY_EARNINGS`) onto a constant `YRSSCH, AgeInDays,` and `AgeInDays` squared. Construct a 95 percent confidence interval for the coefficient on `YrsSch.` Write your own Python function to complete this computation. Your function should also construct and return a variance-covariance estimate which can be used to construct asymptotic standard errors. Compare your results – point estimates and standard errors – with those of the statsmodels OLS implementation.

[b] Compute the least squares fit of ln(`MONTHLY_EARNINGS`) onto a constant `YRSSCH, AgeInDays, AgeInDays` squared, `Dad_NoSchool_c, Dad_1stPrim_c, Dad_2ndPrim_c, Dad_Sec_c, Mom_NoSchool_c, Mom_1stPrim_c, Mom_2ndPrim_c,` and `Mom_Sec_c.` Compare the resulting coefficient on `YRSSCH` with that in part [a] above. Provide an explanation for any differences found.

[c] Show how you can compute the coefficient on `YRSSCH` in (2) by a least squares fit of ln(`MONTHLY_EARNINGS`) on a single variable. Describe this variable, construct it and calculate the least squares fit to check your answer.

[d] Using the Bayes Bootstrap to approximate a posterior distribution of the coefficient on `YRSSCH` in the linear predictors described in parts (a) and (b). How do these posterior distributions compare with their estimated asymptotic sampling distributions?