Exercises

Exercise 3.1 Let y be a random variable with $\mu = \mathbb{E}y$ and $\sigma^2 = \text{var}(y)$. Define

$$g\left(y, \mu, \sigma^2\right) = \begin{pmatrix} y - \mu \\ (y-\mu)^2 - \sigma^2 \end{pmatrix}.$$

Let $(\hat{\mu}, \hat{\sigma}^2)$ be the values such that $\overline{g}_n(\hat{\mu}, \hat{\sigma}^2) = \mathbf{0}$ where $\overline{g}_n(m, s) = n^{-1} \sum_{i=1}^n g\left(y_i, m,$ Show that $\hat{\mu}$ and $\hat{\sigma}^2$ are the sample mean and variance.

Exercise 3.2 Consider the OLS regression of the $n \times 1$ vector $\boldsymbol{y}$ on the $n \times k$ matrix $\boldsymbol{X}$. Consider an alternative set of regressors $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{C}$, where $\boldsymbol{C}$ is a $k \times k$ non-singular matrix. Thus, each column of $\boldsymbol{Z}$ is a mixture of some of the columns of $\boldsymbol{X}$. Compare the OLS estimates and residuals from the regression of $\boldsymbol{y}$ on $\boldsymbol{X}$ to the OLS estimates from the regression of $\boldsymbol{y}$ on $\boldsymbol{Z}$.

Exercise 3.3 Using matrix algebra, show $\boldsymbol{X}'\hat{\boldsymbol{e}} = \mathbf{0}$.

Exercise 3.4 Let $\hat{\boldsymbol{e}}$ be the OLS residual from a regression of $\boldsymbol{y}$ on $\boldsymbol{X} = [\boldsymbol{X}_1$ $\boldsymbol{X}_2]$. Find $\boldsymbol{X}_2'\hat{\boldsymbol{e}}$.

Exercise 3.5 Let $\hat{\boldsymbol{e}}$ be the OLS residual from a regression of $\boldsymbol{y}$ on $\boldsymbol{X}$. Find the OLS coefficient from a regression of $\hat{\boldsymbol{e}}$ on $\boldsymbol{X}$.

Exercise 3.6 Let $\hat{\boldsymbol{y}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$. Find the OLS coefficient from a regression of $\hat{\boldsymbol{y}}$ on $\boldsymbol{X}$.

Exercise 3.7 Show that if $\boldsymbol{X} = [\boldsymbol{X}_1\ \boldsymbol{X}_2]$ then $\boldsymbol{P}\boldsymbol{X}_1 = \boldsymbol{X}_1$ and $\boldsymbol{M}\boldsymbol{X}_1 = \mathbf{0}$.

Exercise 3.8 Show that $\boldsymbol{M}$ is idempotent: $\boldsymbol{M}\boldsymbol{M} = \boldsymbol{M}$.

Exercise 3.9 Show that $\text{tr}\,\boldsymbol{M} = n - k$.

Exercise 3.10 Show that if $\boldsymbol{X} = [\boldsymbol{X}_1\ \boldsymbol{X}_2]$ and $\boldsymbol{X}_1'\boldsymbol{X}_2 = 0$ then $\boldsymbol{P} = \boldsymbol{P}_1 + \boldsymbol{P}_2$.

Exercise 3.11 Show that when $\boldsymbol{X}$ contains a constant, $\frac{1}{n}\sum_{i=1}^{n}\hat{y}_i = \bar{y}$.

Exercise 3.12 A dummy variable takes on only the values 0 and 1. It is used for categorical data, such as an individual's gender. Let $\boldsymbol{d}_1$ and $\boldsymbol{d}_2$ be vectors of 1's and 0's, with the i'th element of $\boldsymbol{d}_1$ equaling 1 and that of $\boldsymbol{d}_2$ equaling 0 if the person is a man, and the reverse if the person is a woman. Suppose that there are n_1 men and n_2 women in the sample. Consider fitting the following three equations by OLS

$$\boldsymbol{y} = \mu + \boldsymbol{d}_1\alpha_1 + \boldsymbol{d}_2\alpha_2 + \boldsymbol{e} \tag{3.46}$$

$$\boldsymbol{y} = \boldsymbol{d}_1\alpha_1 + \boldsymbol{d}_2\alpha_2 + \boldsymbol{e} \tag{3.47}$$

$$\boldsymbol{y} = \mu + \boldsymbol{d}_1\phi + \boldsymbol{e} \tag{3.48}$$

Can all three equations (3.46), (3.47), and (3.48) be estimated by OLS? Explain if not.

(a) Compare regressions (3.47) and (3.48). Is one more general than the other? Explain the relationship between the parameters in (3.47) and (3.48).

(b) Compute $\boldsymbol{\iota}'\boldsymbol{d}_1$ and $\boldsymbol{\iota}'\boldsymbol{d}_2$, where $\boldsymbol{\iota}$ is an $n \times 1$ is a vector of ones.

(c) Letting $\boldsymbol{\alpha} = (\alpha_1 \ \alpha_2)'$, write equation (3.47) as $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\alpha} + e$. Consider the assumption $\mathbb{E}(\boldsymbol{x}_i e_i) = 0$. Is there any content to this assumption in this setting?

Exercise 3.13 Let $\boldsymbol{d}_1$ and $\boldsymbol{d}_2$ be defined as in the previous exercise.

(a) In the OLS regression

$$\boldsymbol{y} = \boldsymbol{d}_1\hat{\gamma}_1 + \boldsymbol{d}_2\hat{\gamma}_2 + \hat{\boldsymbol{u}},$$

show that $\hat{\gamma}_1$ is the sample mean of the dependent variable among the men of the sample ($\overline{y}_1$), and that $\hat{\gamma}_2$ is the sample mean among the women ($\overline{y}_2$).

(b) Let $\boldsymbol{X}$ $(n \times k)$ be an additional matrix of regressions. Describe in words the transformations

$$\boldsymbol{y}^* = \boldsymbol{y} - \boldsymbol{d}_1\overline{y}_1 - \boldsymbol{d}_2\overline{y}_2$$
$$\boldsymbol{X}^* = \boldsymbol{X} - \boldsymbol{d}_1\overline{\boldsymbol{x}}_1' - \boldsymbol{d}_2\overline{\boldsymbol{x}}_2'$$

where $\overline{\boldsymbol{x}}_1$ and $\overline{\boldsymbol{x}}_2$ are the $k \times 1$ means of the regressors for men and women, respectively.

(c) Compare $\widetilde{\boldsymbol{\beta}}$ from the OLS regresion

$$\boldsymbol{y}^* = \boldsymbol{X}^*\widetilde{\boldsymbol{\beta}} + \tilde{\boldsymbol{e}}$$

with $\widehat{\boldsymbol{\beta}}$ from the OLS regression

$$\boldsymbol{y} = \boldsymbol{d}_1\hat{\alpha}_1 + \boldsymbol{d}_2\hat{\alpha}_2 + \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \hat{\boldsymbol{e}}.$$

Exercise 3.14 Let $\widehat{\boldsymbol{\beta}}_n = (\boldsymbol{X}_n'\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n'\boldsymbol{y}_n$ denote the OLS estimate when $\boldsymbol{y}_n$ is $n \times 1$ and $\boldsymbol{X}_n$ is $n \times k$. A new observation $(y_{n+1}, \boldsymbol{x}_{n+1})$ becomes available. Prove that the OLS estimate computed using this additional observation is

$$\widehat{\boldsymbol{\beta}}_{n+1} = \widehat{\boldsymbol{\beta}}_n + \frac{1}{1 + \boldsymbol{x}_{n+1}'(\boldsymbol{X}_n'\boldsymbol{X}_n)^{-1}\boldsymbol{x}_{n+1}}(\boldsymbol{X}_n'\boldsymbol{X}_n)^{-1}\boldsymbol{x}_{n+1}\left(y_{n+1} - \boldsymbol{x}_{n+1}'\widehat{\boldsymbol{\beta}}_n\right).$$

Exercise 3.15 Prove that R^2 is the square of the sample correlation between $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$.

Exercise 3.16 Consider two least-squares regressions

$$\boldsymbol{y} = \boldsymbol{X}_1\widetilde{\boldsymbol{\beta}}_1 + \tilde{\boldsymbol{e}}$$

and

$$\boldsymbol{y} = \boldsymbol{X}_1\widehat{\boldsymbol{\beta}}_1 + \boldsymbol{X}_2\widehat{\boldsymbol{\beta}}_2 + \hat{\boldsymbol{e}}.$$

Let R_1^2 and R_2^2 be the R-squared from the two regressions. Show that $R_2^2 \geq R_1^2$. Is there a case (explain) when there is equality $R_2^2 = R_1^2$?

Exercise 3.17 Show that $\tilde{\sigma}^2 \geq \hat{\sigma}^2$. Is equality possible?

Exercise 3.18 For which observations will $\widehat{\boldsymbol{\beta}}_{(-i)} = \widehat{\boldsymbol{\beta}}$?

Exercise 3.19 Use the data set from Section 3.19 and the sub-sample used for equation (3.43) (see Section 3.20) for data construction)

1. Estimate equation (3.43) and compute the equation R^2 and sum of squared errors.
2. Re-estimate the slope on education using the residual regression approach. Regress log(Wage) on experience and its square, regress education on experience and its square, and the residuals on the residuals. Report the estimates from this final regression, along with the equation R^2 and sum of squared errors. Does the slope coefficient equal the value in (3.43)? Explain.
3. Do the R^2 and sum-of-squared errors from parts 1 and 2 equal? Explain.

Exercise 3.20 Estimate equation (3.43) as in part 1 of the previous question. Let $\hat{e}_i$ be the OLS residual, $\hat{y}_i$ the predicted value from the regression, x_{1i} be education and x_{2i} be experience. Numerically calculate the following:

(a) $\sum_{i=1}^{n} \hat{e}_i$

(b) $\sum_{i=1}^{n} x_{1i}\hat{e}_i$

(c) $\sum_{i=1}^{n} x_{2i}\hat{e}_i$

(d) $\sum_{i=1}^{n} x_{1i}^2\hat{e}_i$

(e) $\sum_{i=1}^{n} x_{2i}^2\hat{e}_i$

(f) $\sum_{i=1}^{n} \hat{y}_i\hat{e}_i$

(g) $\sum_{i=1}^{n} \hat{e}_i^2$

Are these calculations consistent with the theoretical properties of OLS? Explain.

Exercise 3.21 Use the data set from Section 3.19.

1. Estimate a log wage regression for the subsample of white male Hispanics. In addition to education, experience, and its square, include a set of binary variables for regions and marital status. For regions, you create dummy variables for Northeast, South and West so that Midwest is the excluded group. For marital status, create variables for married, windowed or divorced, and separated, so that single (never married) is the excluded group.
2. Repeat this estimation using a different econometric package. Compare your results. Do they agree?