COMM8102, T2 2021

Solutions to Problem Set #1

Q1. Exercise 2.4

$$
\begin{aligned}
E[Y|X=0] =& 0 \times \frac{0.1}{0.5} + 1 \times \frac{0.4}{0.5} = 0.8, \\
E[Y|X=1] =& 0 \times \frac{0.2}{0.5} + 1 \times \frac{0.3}{0.5} = 0.6.
\end{aligned}
$$

Since $Y^2 = Y$ (Y is binary), $E[Y^2|X] = E[Y|X]$. Finally

$$
\begin{aligned}
var(Y|X=0) =& E[Y^2|X=0] - (E[Y|X=0])^2 = E[Y|X=0] - (E[Y|X=0])^2 = 0.8 - 0.64 = 0.16, \\
var(Y|X=1) =& E[Y^2|X=1] - (E[Y|X=1])^2 = E[Y|X=1] - (E[Y|X=1])^2 = 0.6 - 0.36 = 0.24.
\end{aligned}
$$

Q2. Exercise 2.7

The conditional variance of Y given X is $E[(Y - E[Y|X])^2|X]$ by definition. This definition can be written as

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

as desired.

Q3. Exercise 2.18

(a) Assume that $\alpha_1 \neq 0$ or $\alpha_2 \neq 0$. To show $\mathbf{Q}_{XX}$ is not invertible, we show that the columns are linearly dependent (i.e., one column can be written as a non-zero linear combination of other columns). Note that there are other ways to shows this (e.g. determinant). Since $X_3 = \alpha_1 + \alpha_2 X_2$,

$$
\begin{aligned}
\mathbf{Q}_{XX} = \begin{pmatrix} 1 \\ X_2 \\ X_3 \end{pmatrix} \begin{pmatrix} 1 & X_2 & X_3 \end{pmatrix} =& \begin{pmatrix} 1 & X_2 & X_3 \\ X_2 & X_2^2 & X_2 X_3 \\ X_3 & X_2 X_3 & X_3^2 \end{pmatrix} \\
=& \begin{pmatrix} 1 & X_2 & \alpha_1 + \alpha_2 X_2 \\ X_2 & X_2^2 & X_2(\alpha_1 + \alpha_2 X_2) \\ \alpha_1 + \alpha_2 X_2 & X_2(\alpha_1 + \alpha_2 X_2) & (\alpha_1 + \alpha_2 X_2)^2 \end{pmatrix}.
\end{aligned}
$$

Let q_j be the jth column of the matrix $\mathbf{Q}_{XX}$. Then $q_3 = \alpha_1 q_1 + \alpha_2 q_2$. Thus, $\mathbf{Q}_{XX}$ is not invertible.

(b) The best linear predictor of Y given X is $X'\beta$ where $\beta = (E[XX'])^{-1}E[XY]$ is the projection coefficient. But we cannot calculate β because $\mathbf{Q}_{XX}$ is not invertible. We can use the fact that the linear space spanned by $X^* = (1, X_2)'$ is exactly same with the one spanned by X because $X_3 = \alpha_1 + \alpha_2 X_2$. Therefore, the best linear predictor of Y given X is identical to the best linear predictor of Y given X^*, which is $X^{*'}\beta^*$ where

$$\begin{aligned}
X^{*'}\beta^* =& X^{*'}(E[X^*X^{*'}])^{-1}E[X^*Y] \\
=& \begin{pmatrix} 1 & X_2 \end{pmatrix} \begin{pmatrix} 1 & E[X_2] \\ E[X_2] & E[X_2^2] \end{pmatrix}^{-1} \begin{pmatrix} E[Y] \\ E[X_2Y] \end{pmatrix} \\
=& \begin{pmatrix} 1 & X_2 \end{pmatrix} \frac{1}{var(X_2)} \begin{pmatrix} E[X_2^2] & -E[X_2] \\ -E[X_2] & 1 \end{pmatrix} \begin{pmatrix} E[Y] \\ E[X_2Y] \end{pmatrix} \\
=& \begin{pmatrix} 1 & X_2 \end{pmatrix} \frac{1}{var(X_2)} \begin{pmatrix} E[X_2^2]E[Y] - E[X_2]E[X_2Y] \\ cov(X_2, Y) \end{pmatrix} \\
=& \frac{1}{var(X_2)} \left(E[X_2^2]E[Y] - E[X_2]E[X_2Y] + X_2 cov(X_2, Y) \right).
\end{aligned}$$

Q4. Exercise 3.3

$$\mathbf{X}'\widehat{\mathbf{e}} = \mathbf{X}'(\mathbf{Y} - \widehat{\mathbf{Y}}) = \mathbf{X}'(\mathbf{Y} - \mathbf{X}\widehat{\beta}) = \mathbf{X}'(\mathbf{Y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\mathbf{Y}) = \mathbf{0}.$$

Q5. Exercise 3.12

(3.53) and (3.54) can be estimated by OLS, but not (3.52) because of multicollinearity.

(a) (3.53) and (3.54) are observationally equivalent (i.e., one is not more general than the other):

$$\begin{aligned}
\mathbf{Y} &= \mathbf{D}_1\alpha_1 + \mathbf{D}_2\alpha_2 + \mathbf{e} \\
&= \mathbf{D}_1\alpha_1 + (1 - \mathbf{D}_1)\alpha_2 + \mathbf{e} \\
&= \alpha_2 + \mathbf{D}_1(\alpha_1 - \alpha_2) + \mathbf{e}.
\end{aligned}$$

Thus, $\mu = \alpha_2$, $\phi = \alpha_1 - \alpha_2$.

(b) $\iota'\mathbf{D}_1 = n_1$, $\iota'\mathbf{D}_2 = n_2$.

Q6. Exercise 3.19

Using the definition of the leave-one-out (LOO) estimator, we derive

$$\widehat{\beta}_{(-i)} = \frac{1}{n-1}\sum_{j \neq i} Y_j$$

where $\sum_{j \neq i}$ is summation over all $j = 1, ..., n$ except for $j = i$. The LOO predicted value is

$$\widetilde{Y}_i = \widehat{\beta}_{(-i)}$$

and the LOO prediction error is

$$\widetilde{e}_i = Y_i - \widetilde{Y}_i = Y_i - \widehat{\beta}_{(-i)}.$$

But we can write

$$\begin{aligned}
\widetilde{e}_i =& Y_i - \frac{1}{n-1} \sum_{j \neq i} Y_j \\
=& Y_i - \frac{n}{n-1} \frac{1}{n} \left(\sum_{j=1}^{n} Y_j - Y_i \right) \\
=& Y_i - \frac{n}{n-1} \left(\overline{Y} - \frac{1}{n} Y_i \right) \\
=& \left(1 + \frac{1}{n-1} \right) Y_i - \frac{n}{n-1} \overline{Y} \\
=& \frac{n}{n-1} \left(Y_i - \overline{Y} \right),
\end{aligned}$$

as desired.

Q7. Exercise 3.24 (2 points)
The textbook equation (3.49) is the regression of log wage on education, experience, experience squared, and a constant using the subsample of individuals who are single Asian men with less than 45 years experience.
(a) We can replicate the equation (3.49). R^2 and the sum of squared errors are 0.3893 and 82.5050, respectively.
(b) By the FWL theorem, the estimated coefficient for *education* is identical to (a). R^2 and the sum of squared errors are 0.3687 and 82.5050, respectively.
(c) Let $Y_i = log(wage_i)$. Recall the formula for R^2:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \widehat{e}_i^2}{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}. \tag{1}$$

For the residual regression case, the denominator is $\sum_{i=1}^{n} \widetilde{e}_i^2$, where $\widetilde{e}_i$ be the residual after regressing Y_i on experience, its square, and the constant. Since the analysis-of-variance formula is

$$\sum_{i=1}^{n} (Y_i - \overline{Y})^2 = \sum_{i=1}^{n} (\widetilde{Y}_i - \overline{Y})^2 + \sum_{i=1}^{n} \widetilde{e}_i^2, \tag{2}$$

where $\widetilde{Y}_i$ is the fitted value of the regression of Y_i on experience, its square, and the constant, we have

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 \geq \sum_{i=1}^{n}\tilde{e}_i^2. \tag{3}$$

As a result, the R^2 from the original regression is larger than the one from the residual regression. Note that if you include the constant, then you don't have to include in the final regression, but otherwise, the constant should be included.

Q8. Exercise 3.26 (a) (2 points)

Using the subsample of white male Hispanics, a log wage regression on education, experience and its square, and dummy variables for regions and marital status is estimated. The number of subsample is 4230. The estimated equation is

$$\begin{aligned} \log(wage_i) &= \beta_1 \cdot education + \beta_2 \cdot experience + \beta_3 \cdot experience^2/100 \\ &\quad + \beta_4 \cdot northeast + \beta_5 \cdot south + \beta_6 \cdot west \\ &\quad + \beta_7 \cdot married + \beta_8 \cdot widowed \text{ or } divorced + \beta_{10} \cdot separated + \alpha. \end{aligned} \tag{4}$$

The estimation results are as follows.

log(wage)	$\hat{\beta}$
Education	0.0883
Experience	0.0279
Experience^2/100	-0.0363
Northeast	0.0616
South	-0.0679
West	0.0195
Married	0.1778
Widowed or Divorced	0.0857
Separated	0.0167
Intercept	1.1929

Table 1: Linear regression coefficients of log wage equations without interaction terms