Statistical Learning

Homework 2

Part 1: Conceptual Questions

Problem 1. In an econometric model, we say that a parameter is identified if we can recover its value perfectly given the joint distribution of the observable variables. Suppose that (Y, X) is the observable variables and U is the unobservable variable.

1. Suppose that $Y = \beta_0 + \beta_1 X + U$ and $\mathrm{E}[U] = \mathrm{E}[XU] = 0$. Show that β_1 is identified. I.e., if you know the joint distribution of (Y, X), how do you determine the value of the parameter β_1?

2. Suppose that Y is binary and $Y = 1(\beta_0 + \beta_1 X \geq U)$ and U is a standard normal $(\mathrm{N}(0,1))$ random variable that is independent of X. If you know the joint distribution of (Y, X), how do you determine the value of the parameter β_1? Hint: $\mathrm{E}[Y \mid X] = \mathrm{E}[1(\beta_0 + \beta_1 X \geq U) \mid X] = \Phi(\beta_0 + \beta_1 X)$, where Φ is the standard normal CDF.

Solution. Take

$$\begin{aligned}\mathrm{Cov}[Y, X] = \mathrm{Cov}[\beta_0 + \beta_1 X + U, X] = \mathrm{Cov}[\beta_1 X + U, X] = \\ \beta_1 \mathrm{Cov}[X, X] + \mathrm{Cov}[U, X] = \beta_1 \mathrm{Var}[X].\end{aligned}$$

Therefore, $\beta_1 = \mathrm{Cov}[Y, X] / \mathrm{Var}[X]$. This quantity can be recovered if you know the joint distribution of (Y, X).
Similarly, $\mathrm{E}[Y \mid X] = \Phi(\beta_0 + \beta_1 X)$ gives $\beta_0 + \beta_1 X = \Phi^{-1}(\mathrm{E}[Y \mid X])$, where Φ^{-1} is the inverse function of Φ (Φ is strictly increasing). Then,

$$\mathrm{Cov}\left[\Phi^{-1}(\mathrm{E}[Y \mid X]), X\right] = \mathrm{Cov}[\beta_0 + \beta_1 X, X] = \beta_1 \mathrm{Var}[X].$$

Therefore, $\beta_1 = \mathrm{Cov}[\Phi^{-1}(\mathrm{E}[Y \mid X]), X] / \mathrm{Var}[X]$. This quantity can be recovered if you know the joint distribution of (Y, X).

Problem 2. In this question, we show that in linear regression R^2 is a non-decreasing function of the number of the regressors. Consider the sample $(Y_i, X_{1,i}, X_{2,i})$, $i = 1, 2, ..., n$, with two predictors $X_{1,i}, X_{2,i}$. Let $\tilde{\beta}_0, \tilde{\beta}_1$ denote the OLS coefficients of the linear regression of Y_i against $X_{1,i}$. Let $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ denote the OLS coefficients of the linear regression of Y_i against $X_{1,i}, X_{2,i}$. Let $\tilde{U}_i$ and $\hat{U}_i$ denote the OLS residuals respectively. I.e.,

$$\begin{aligned} Y_i &= \tilde{\beta}_0 + \tilde{\beta}_1 X_{1,i} + \tilde{U}_i, \\ Y_i &= \hat{\beta}_0 + \hat{\beta}_1 X_{1,i} + \hat{\beta}_2 X_{2,i} + \hat{U}_i. \end{aligned}$$

1. Show that $\sum_{i=1}^n \tilde{U}_i = \sum_{i=1}^n \tilde{U}_i X_{1,i} = 0$ and $\sum_{i=1}^n \hat{U}_i = \sum_{i=1}^n \hat{U}_i X_{1,i} = \sum_{i=1}^n \hat{U}_i X_{2,i} = 0$.

2. Show that $\sum_{i=1}^{n} \tilde{U}_i \hat{U}_i = \sum_{i=1}^{n} \hat{U}_i^2$.

3. Show that $\sum_{i=1}^{n} \tilde{U}_i^2 \geq \sum_{i=1}^{n} \hat{U}_i^2$.

4. Show that the R^2 from the second (long) regression is larger than that of the first (short) regression.

Solution.

1. By definition, $\left(\tilde{\beta}_0, \tilde{\beta}_1\right)$ minimizes $\sum_{i=1}^{n} (Y_i - b_0 - b_1 X_{1,i})^2$ over (b_0, b_1) and $\left(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2\right)$ minimizes $\sum_{i=1}^{n} (Y_i - b_0 - b_1 X_{1,i} - b_2 X_{2,i})^2$ over (b_0, b_1, b_2). The first-order conditions are satisfied:

$$\begin{aligned} \sum_{i=1}^{n} \left(Y_i - \tilde{\beta}_0 - \tilde{\beta}_1 X_{1,i}\right) &= 0 \\ \sum_{i=1}^{n} \left(Y_i - \tilde{\beta}_0 - \tilde{\beta}_1 X_{1,i}\right) X_{1,i} &= 0 \end{aligned}$$

and

$$\begin{aligned} \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1,i} - \hat{\beta}_2 X_{2,i}\right) &= 0 \\ \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1,i} - \hat{\beta}_2 X_{2,i}\right) X_{1,i} &= 0 \\ \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1,i} - \hat{\beta}_2 X_{2,i}\right) X_{2,i} &= 0. \end{aligned}$$

2. By Part 1,

$$\begin{aligned} \sum_{i=1}^{n} \tilde{U}_i \hat{U}_i &= \sum_{i=1}^{n} \left(Y_i - \tilde{\beta}_0 - \tilde{\beta}_1 X_{1,i}\right) \hat{U}_i \\ &= \sum_{i=1}^{n} Y_i \hat{U}_i \\ &= \sum_{i=1}^{n} \left(\hat{\beta}_0 + \hat{\beta}_1 X_{1,i} + \hat{\beta}_2 X_{2,i} + \hat{U}_i\right) \hat{U}_i \\ &= \sum_{i=1}^{n} \hat{U}_i^2. \end{aligned}$$

3. By Part 2,

$$0 \leq \sum_{i=1}^{n} \left(\tilde{U}_i - \hat{U}_i\right)^2 = \sum_{i=1}^{n} \tilde{U}_i^2 + \sum_{i=1}^{n} \hat{U}_i^2 - 2 \sum_{i=1}^{n} \tilde{U}_i \hat{U}_i = \sum_{i=1}^{n} \tilde{U}_i^2 - \sum_{i=1}^{n} \hat{U}_i^2.$$

4. Let R_{ur}^2 denote the R^2 from the long regression. Let R_r^2 denote the R^2 from the short regression. Then, $R_{ur}^2 = 1-\sum_{i=1}^n \hat{U}_i^2 / \sum_{i=1}^n \left(Y_i - \overline{Y}\right)^2$ and $R_r^2 = 1-\sum_{i=1}^n \tilde{U}_i^2 / \sum_{i=1}^n \left(Y_i - \overline{Y}\right)^2$, where $\overline{Y} = n^{-1} \sum_{i=1}^n Y_i$. Clearly, $R_{ur}^2 \geq R_r^2$.

Problem 3. Question 4 on Page 189 (ISL second edition).

Solution.

1. If $x \in [0.05, 0.95]$, then the observations used for prediction are in the interval $[x - 0.05, x + 0.05]$. If $x < 0.05$, the observations used for prediction in the interval $[0, x + 0.05]$ which represents a fraction of $(100x + 5)\,\%$. If $x > 0.95$, then the fraction of observations is $(105 - 100x)\,\%$. To compute the average fraction,

$$\int_{0.05}^{0.95} 10dx + \int_{0}^{0.05} (100x + 5)\, dx + \int_{0.95}^{1} (105 - 100x)\, dx = 9 + 0.375 + 0.375 = 9.75.$$

   On average, the fraction of observations for prediction is 9.75%.

2. If we assume X_1 and X_2 to be independent, the fraction of observations for prediction is $9.75\%^2 \approx 0.95\%$.

3. The fraction of observations for prediction is $(9.75\%)^{100} \approx 0$.

4. The fraction of observations for prediction is $(9.75\%)^p$. We have $\lim_{p\uparrow\infty} (9.75\%)^p = 0$.

5. Let ℓ denote the length of the cube. For $p = 1$, $\ell = 0.1$. For $p = 2$, $\ell^2 = 0.1$. For $p = 100$, $\ell^{100} = 0.1$.

Problem 4. Define a density function

$$f(x \mid \theta) = \begin{cases} \left(1 + \frac{1-2\theta}{\theta-1}\right) x^{\frac{1-2\theta}{\theta-1}} & x \in (0,1) \\ 0 & x \notin (0,1), \end{cases}$$

where $0 < \theta < 1$ is a parameter. $X_1, ..., X_n$ is an independent and identically distributed sample with true density $f(\cdot \mid \theta_*)$ for some θ_*.

1. Show that $f(\cdot \mid \theta)$ is a probability density function, for all $0 < \theta < 1$.

2. Show that $\theta_* = \int_0^1 x f(x \mid \theta_*)\, \mathrm{d}x$. I.e., in this parametrization, θ_* is also the population mean. Derive the method of moment estimator of θ_*.

3. Write the log-maximum likelihood function and derive the the maximum likelihood estimator. Is it equal to the method of moment estimator?

Solution.

1. Compute

$$\int_0^1 f(x \mid \theta)\, dx = \left(1 + \frac{1-2\theta}{\theta-1}\right) \int_0^1 x^{\frac{1-2\theta}{\theta-1}}\, dx = \left(1 + \frac{1-2\theta}{\theta-1}\right) \frac{1}{1 + \frac{1-2\theta}{\theta-1}} x^{1+\frac{1-2\theta}{\theta-1}} \Bigg|_0^1 = 1.$$

   Therefore, $f(x \mid \theta) \geq 0$ and $\int_0^1 f(x \mid \theta)\, dx = 1$.

2. Compute

$$\int_0^1 xf(x \mid \theta_*)\,dx = \left(1+\frac{1-2\theta_*}{\theta_*-1}\right)\int_0^1 x\cdot x^{\frac{1-2\theta_*}{\theta_*-1}}\,dx = \left(1+\frac{1-2\theta_*}{\theta_*-1}\right)\frac{1}{1-\frac{\theta_*}{\theta_*-1}}x^{1-\frac{\theta_*}{\theta_*-1}}\Big|_0^1 = \theta_*.$$

The method of moment estimator: $n^{-1}\sum_{i=1}^n X_i$.

- The log-maximum likelihood function is

$$\log L(\theta; X_1, ..., X_n) = n\log\left(\frac{\theta}{1-\theta}\right) + \frac{1-2\theta}{\theta-1}\sum_{i=1}^n \log(X_i).$$

Differentiating with respect to θ:

$$\frac{\partial \log L}{\partial\theta} = \frac{n}{\theta(1-\theta)} + \frac{1}{(1-\theta)^2}\sum_{i=1}^n \log(X_i).$$

Solving the first order condition, the maximum likelihood estimator is

$$\hat{\theta} = \frac{n}{n-\sum_{i=1}^n \log(X_i)},$$

which is different from the method of moments estimator.

Problem 5. Given training data $\mathsf{Tr} = \{(X_1, Y_1), ..., (X_n, Y_n)\}$, suppose that $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ where ϵ_i is the error term. Denote $X_1^n = (X_1, ..., X_n)$ for notational simplicity. Assume that $\mathrm{E}[\epsilon_i \mid X_1^n] = 0$, $\mathrm{E}[\epsilon_i^2 \mid X_1^n] = \sigma^2$ (for some $\sigma^2 > 0$) and $\mathrm{E}[\epsilon_i\epsilon_j \mid X_1^n] = 0$, $\forall i$ and $\forall j \neq i$. Assume that the conditional distribution of ϵ_i given X_1^n is $\mathrm{N}(0, \sigma^2)$. Let $\hat{\beta}_0, \hat{\beta}_1$ denote the OLS estimator. Let x_0 be a fixed value and $y_0 = \beta_0 + \beta_1 x_0$. Let $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ be the estimator of y_0. Let $Y_0 = y_0 + \epsilon_0$, where ϵ_0 denotes an error that is independent of the training data Tr ($\epsilon_0 \mid \mathsf{Tr} \sim \mathrm{N}(0, \sigma^2)$). In this question, assume that σ^2 is known.

1. Show that $\mathrm{E}[\hat{y}_0 \mid X_1^n] = y_0$ find the expression of $\mathrm{Var}[\hat{y}_0 \mid X_1^n]$.

2. What is conditional distribution of $\hat{y}_0$ given X_1^n?

3. What is conditional variance of $\hat{y}_0 - Y_0$ given X_1^n? Hint: $\mathrm{E}[\epsilon_0 \mid \mathsf{Tr}] = \mathrm{E}[\epsilon_0] = 0$ and by law of iterated expectations,

$$\mathrm{E}[\epsilon_0\hat{y}_0 \mid X_1^n] = \mathrm{E}[\mathrm{E}[\epsilon_0\hat{y}_0 \mid \mathsf{Tr}]] = \mathrm{E}[\hat{y}_0\mathrm{E}[\epsilon_0 \mid \mathsf{Tr}]] = 0.$$

What is conditional distribution of $\hat{y}_0 - Y_0$ given X_1^n?

4. Propose a prediction interval $[LB, UB]$ that covers Y_0 with probability 95%. Find LB and UB.

Solution.

1. Denote $\overline{Y} = n^{-1}\sum_{i=1}^{n} Y_i$, $\overline{X} = n^{-1}\sum_{i=1}^{n} X_i$ and $\overline{\epsilon} = n^{-1}\sum_{i=1}^{n} \epsilon_i$. We have $\hat{\beta}_1 = \sum_{i=1}^{n}\left(X_i - \overline{X}\right)Y_i / \sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$, $\hat{\beta}_0 = \overline{Y} - \overline{X}\hat{\beta}_1$ and $\overline{Y} = \beta_0 + \beta_1\overline{X} + \overline{\epsilon}$. Then, $\hat{\beta}_0 = \beta_0 + \beta_1\overline{X} + \overline{\epsilon} - \overline{X}\hat{\beta}_1 = \beta_0 + \overline{\epsilon} - \overline{X}\left(\hat{\beta}_1 - \beta_1\right)$. And,

$$\begin{aligned}\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 &= \beta_0 + \overline{\epsilon} - \overline{X}\left(\hat{\beta}_1 - \beta_1\right) + \hat{\beta}_1 x_0 \\ &= \beta_0 + \beta_1 x_0 + \overline{\epsilon} - \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(\overline{X} - x_0\right)\epsilon_i}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} \\ &= \beta_0 + \beta_1 x_0 + \frac{1}{n}\sum_{i=1}^{n}\left\{1 - \frac{\left(X_i - \overline{X}\right)\left(\overline{X} - x_0\right)}{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\right\}\epsilon_i.\end{aligned}$$

It follows that $\mathrm{E}\left[\hat{y}_0 \mid X_1^n\right] = y_0$ and

$$\mathrm{Var}\left[\hat{y}_0 \mid X_1^n\right] = \frac{1}{n^2}\sum_{i=1}^{n}\left\{1 - \frac{\left(X_i - \overline{X}\right)\left(\overline{X} - x_0\right)}{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\right\}^2 \sigma^2 = \frac{1}{n}\left\{1 + \frac{\left(\overline{X} - x_0\right)^2}{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\right\}\sigma^2.$$

2. Conditional on X_1^n, $\hat{y}_0$ is a linear function of $(\epsilon_1, \epsilon_2, ..., \epsilon_n)$, which is jointly normal. Therefore, $\hat{y}_0 \mid X_1^n \sim \mathrm{N}\left(y_0, \mathrm{Var}\left[\hat{y}_0 \mid X_1^n\right]\right)$.

3. $\hat{y}_0 - Y_0 = \hat{y}_0 - y_0 - \epsilon_0$ and ϵ_0 is independent of $\hat{y}_0 - y_0$. Then,

$$\mathrm{Var}\left[\hat{y}_0 - Y_0 \mid X_1^n\right] = \mathrm{E}\left[\left(\hat{y}_0 - y_0 - \epsilon_0\right)^2 \mid X_1^n\right] = \frac{1}{n}\left\{1 + \frac{\left(\overline{X} - x_0\right)^2}{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\right\}\sigma^2 + \sigma^2,$$

since $\mathrm{E}\left[\epsilon_0\left(\hat{y}_0 - y_0\right) \mid X_1^n\right] = 0$. $\hat{y}_0 - Y_0$ is a linear function of $(\epsilon_1, \epsilon_2, ..., \epsilon_n, \epsilon_0)$. $\hat{y}_0 - Y_0 \mid X_1^n \sim \mathrm{N}\left(\mathrm{E}\left[\hat{y}_0 - Y_0 \mid X_1^n\right], \mathrm{Var}\left[\hat{y}_0 - Y_0 \mid X_1^n\right]\right)$. And, it is easy to check $\mathrm{E}\left[\hat{y}_0 - Y_0 \mid X_1^n\right] = 0$.

4. Since

$$\hat{y}_0 - Y_0 \mid X_1^n \sim \mathrm{N}\left(0, \frac{1}{n}\left\{1 + \frac{\left(\overline{X} - x_0\right)^2}{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\right\}\sigma^2 + \sigma^2\right)$$

and therefore,

$$\frac{\hat{y}_0 - Y_0}{SE} \sim \mathrm{N}(0,1), \text{ with } SE = \sqrt{\frac{1}{n}\left\{1 + \frac{\left(\overline{X} - x_0\right)^2}{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\right\}\sigma^2 + \sigma^2}.$$

Then, $\Pr\left[\left|\left(\hat{y}_0 - Y_0\right)/SE\right| \leq 1.96\right] = 95\%$. And therefore, $LB = \hat{y}_0 - 1.96 \cdot SE$ and $UB = \hat{y}_0 + 1.96 \cdot SE$.

Part 2: Applied Questions

Problem 6. Question 8 on Page 123 (ISL second edition).

Problem 7. Question 9 on Page 123 (ISL second edition).

Problem 8. Question 13 on Page 193 (ISL second edition).