24500 HW8

Bin Yu

March 03, 2025

Question 1

(a)

In this part, we assume the model

$$y_i \sim N(\beta_0, 1)$$

with a true value $\beta_0 = 3$. We generate $n = 30$ independent observations from $N(3,1)$ and construct the 95% confidence interval for β_0 using the sample mean $\bar{y}$. Specifically, we use the formula

$$\bar{y} \;\pm\; z_{1-\alpha/2}\,\frac{1}{\sqrt{n}},$$

where $z_{1-\alpha/2}$ is the critical value of the standard normal distribution for the $(1-\alpha/2)$ quantile (approximately 1.96 for $\alpha = 0.05$).

With $\alpha = 0.05$, we repeat this procedure for many simulations (10^5 times) and record how often $\beta_0 = 3$ is covered.

R Output

```
coverage_rate <- mean(coverage)
print(coverage_rate)
[1] 0.95052
```

By repeating this simulation experiment 10^5 times, we count the proportion of intervals that contain the true mean $\beta_0 = 3$. The empirical coverage rate is about 0.95052, which is very close to the nominal 95%, confirming that this confidence interval achieves the intended coverage.

(b)

According to the problem statement, we can construct a 95% confidence interval for β_0 by first computing the ordinary least squares (OLS) estimates $\hat{\beta}_0, \hat{\beta}_1$. Under the assumption $\sigma^2 = 1$ is known, the formula for the 95% confidence interval is

$$\hat{\beta}_0 \;\pm\; z_{1-\alpha/2}\sqrt{\frac{1}{n} \;+\; \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}},$$

We will assume the data come from the linear model

$$y_i \;=\; \beta_0 + \beta_1\,x_i \;+\; \varepsilon_i, \quad \varepsilon_i \sim N(0,1).$$

Here we have chosen $\beta_0 = 3$ and $\beta_1 = 0$, because we generate $y_i \sim N(3,1)$ (i.e. a constant mean of 3 that does not depend on x_i), and $x_i = \frac{i}{10}$.

With $\alpha = 0.05$, we repeat this procedure for many simulations (10^5 times) and record how often $\beta_0 = 3$ is covered.

R Output

```
coverage_rate <- mean(coverage)
print(coverage_rate)
# output:
# [1] 0.95035
```

By repeating this simulation experiment 10^5 times, the empirical coverage rate is approximately 0.95035, which is very close to the nominal 95%. Note that here the true parameters are effectively $\beta_0 = 3$ and $\beta_1 = 0$, since the data generation $y_i \sim N(3, 1)$ does not depend on x_i.

(c)

We are uncertain whether the true model is a constant model

$$y_i \sim N(\beta_0, 1)$$

or a linear model

$$y_i \sim N(\beta_0 + \beta_1 x_i, 1).$$

Hence, we first compute the ordinary least squares (OLS) estimators $\hat{\beta}_0, \hat{\beta}_1$. Then:

$$\text{If } \left|\hat{\beta}_1\right| \;\leq\; z_{1-\alpha/2}\sqrt{\frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}},$$

we conclude the slope is "not significant" and choose the constant model's confidence interval (CI). Otherwise, we choose the linear model's CI.

Here we still chose $\beta_0 = 3$ and $\beta_1 = 0$, because we generate $y_i \sim N(3, 1)$ (i.e. a constant mean of 3 that does not depend on x_i).

With $\alpha = 0.05$, we repeat this procedure for many simulations (10^5 times) and record how often $\beta_0 = 3$ is covered.

R Output

```
coverage_rate <- mean(coverage_results)
print(coverage_rate)
# Output
[1] 0.92597
```

By repeating this procedure 10^5 times, we find the empirical coverage of $\beta_0 = 3$. The empirical coverage rate for β_0 in this simulation is approximately 0.92597, which is lower than 95%.

(d)

We again consider the same two possible models:

$$y_i \sim N(\beta_0, 1) \quad \text{vs.} \quad y_i \sim N(\beta_0 + \beta_1 x_i, 1).$$

However, instead of using the same data to both test whether $\beta_1 = 0$ and construct the confidence interval for β_0, we split the sample in half. Concretely, we take a "test set" of size $n/2$ and an "estimation set" of size $n/2$. The procedure can be written as follows:

1. Generate data. We set $\beta_0 = 3$ and $\beta_1 = 0$, so

$$y_i \;=\; 3 + \varepsilon_i, \quad \varepsilon_i \sim N(0,1),$$

   and $x_i = i/10$ for $i = 1, \ldots, n$. In other words, each y_i is drawn from $N(3,1)$, independent of x_i.

2. Test set (first half). We fit the linear model on the first $n/2$ observations to obtain $\hat{\beta}_0^{(\text{test})}$ and $\hat{\beta}_1^{(\text{test})}$. We then perform the test

$$\text{If } \left|\hat{\beta}_1^{(\text{test})}\right| \;\le\; z_{1-\alpha/2}\sqrt{\frac{1}{\sum_{i=1}^{n/2}(x_i - \bar{x})^2}}, \quad \text{choose constant model;} \quad \text{else choose linear model.}$$

   Here, $\alpha = 0.05$.

3. Estimation set (second half). Based on the test result, we then construct the confidence interval for β_0 *only* using the second half of the data:

$$\text{CI}(\beta_0) \;=\; \begin{cases} \bar{y}_{\text{est}} \;\pm\; z_{1-\alpha/2}\,\frac{1}{\sqrt{n/2}}, & \text{if the constant model is selected,} \\ \hat{\beta}_0^{(\text{est})} \;\pm\; z_{1-\alpha/2}\sqrt{\frac{1}{n/2} + \frac{\bar{x}_{\text{est}}^2}{\sum(x_i - \bar{x}_{\text{est}})^2}}, & \text{if the linear model is selected.} \end{cases}$$

   Here, $\bar{y}_{\text{est}}$ and $\hat{\beta}_0^{(\text{est})}$ are computed solely from the estimation subset.

4. Repeat. We repeat this procedure (data generation, test set decision, estimation set CI) for many simulations (10^5 times) and record how often $\beta_0 = 3$ is covered.

R Output

```
cat("Overall coverage:      ", overall_coverage,   "\n")
cat("Coverage (constant):   ", coverage_const,    "\n")
cat("Coverage (linear):     ", coverage_linear,   "\n")
cat("Prop chosen constant:  ", prop_const,        "\n")
cat("Prop chosen linear:    ", prop_linear,       "\n")

Overall coverage:       0.94992
Coverage (constant):    0.9498347
Coverage (linear):      0.9515357
Prop chosen constant:   0.94986
Prop chosen linear:     0.05014
```

By repeating this procedure 10^5 times,

- Overall coverage ≈ 0.94992. By splitting the data, the final procedure achieves an empirical coverage close to the nominal 95%.
- Coverage (constant) ≈ 0.94983, coverage (linear) ≈ 0.95154. Conditioned on which model is chosen, the coverage are still around 95%.

Discussion

For (a) Constant Model $\big(y_i \sim N(\beta_0, 1)\big)$**.**

Suppose $y_1, y_2, \ldots, y_n$ are i.i.d. observations from $N(\beta_0, 1)$. Then the sample mean is

$$\bar{y} \;=\; \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Because the y_i are normally distributed with mean β_0 and variance 1, $\bar{y}$ itself follows

$$\bar{y} \;\sim\; N\Big(\beta_0,\; \frac{1}{n}\Big).$$

Hence the standardized quantity

$$Z \;=\; \sqrt{n}\,\big(\bar{y}-\beta_0\big) \sim N(0,1)$$

For a 95% confidence interval, we know that

$$P\Big(-z_{1-\alpha/2} \;\le\; Z \;\le\; z_{1-\alpha/2}\Big) \;=\; 1-\alpha,$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of the standard normal (approximately 1.96 for $\alpha = 0.05$). Substituting $Z = \sqrt{n}(\bar{y}-\beta_0)$ gives

$$P\Big(-z_{1-\alpha/2} \;\le\; \sqrt{n}\,(\bar{y}-\beta_0) \;\le\; z_{1-\alpha/2}\Big) \;=\; 1-\alpha,$$

$$P\Big(\beta_0 \;\in\; \bar{y} \;\pm\; z_{1-\alpha/2}\,\tfrac{1}{\sqrt{n}}\Big) \;=\; 1-\alpha.$$

Thus the interval

$$\bar{y} \;\pm\; z_{1-\alpha/2}\,\frac{1}{\sqrt{n}}$$

is a valid $(1-\alpha)\times 100\%$ (e.g. 95%) confidence interval for β_0.

For (b) Linear Model $\big(y_i \sim N(\beta_0+\beta_1 x_i, 1)\big)$.

Now let $y_1, y_2, \ldots, y_n$ be i.i.d. observations such that

$$y_i \;=\; \beta_0 + \beta_1\,x_i \;+\; \varepsilon_i, \quad \varepsilon_i \sim N(0,1) \quad \text{i.i.d.}$$

Here β_0 and β_1 are unknown parameters, and $\sigma^2 = 1$ is known. The ordinary least squares (OLS) estimator for β_0 can be shown (via linear regression formulas) to be unbiased with

$$\mathrm{E}[\hat{\beta}_0] \;=\; \beta_0,$$

and when $\sigma^2 = 1$ is known, its variance is

$$\mathrm{Var}(\hat{\beta}_0) \;=\; \frac{1}{n} \;+\; \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2},$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

Hence the standardized variable

$$Z \;=\; \frac{\hat{\beta}_0-\beta_0}{\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{\sum(x_i-\bar{x})^2}}} \sim N(0,1)s$$

By the same reasoning as in part (a), for a 95% confidence interval,

$$P\Big(-z_{1-\alpha/2} \;\le\; Z \;\le\; z_{1-\alpha/2}\Big) \;=\; 1-\alpha,$$

which is equivalent to

$$P\Big(\beta_0 \;\in\; \hat{\beta}_0 \;\pm\; z_{1-\alpha/2}\sqrt{\frac{1}{n} \;+\; \frac{\bar{x}^2}{\sum(x_i-\bar{x})^2}}\Big) \;=\; 1-\alpha.$$

Therefore, the interval

$$\hat{\beta}_0 \;\pm\; z_{1-\alpha/2}\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

is also a valid $(1-\alpha)\times 100\%$ confidence interval for β_0 under the linear model assumption.

For (c)

Let $\widehat{\beta}_1$ be the OLS estimator of β_1, and suppose we choose between two confidence intervals for β_0 based on the event

$$T \;=\; \Big\{\, |\widehat{\beta}_1| \;<\; c \Big\},$$

where c is some critical value derived from a test for $\beta_1 = 0$. Now define two candidate intervals for β_0:

$$\mathrm{CI}_0 = \bar{y} \;\pm\; z_{1-\alpha/2}\,\frac{1}{\sqrt{n}} \quad (\text{if } |\widehat{\beta}_1| < c, \text{ i.e. } T \text{ occurs}),$$

$$\mathrm{CI}_1 = \hat{\beta}_0 \;\pm\; z_{1-\alpha/2}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \quad (\text{if } |\widehat{\beta}_1| \geq c, \text{ i.e. } T^c \text{ occurs}).$$

Method (c) then selects

$$\mathrm{CI}_{\text{selected}} \;=\; \begin{cases} \mathrm{CI}_0, & \text{if } T \text{ is true,} \\ \mathrm{CI}_1, & \text{if } T^c \text{ is true.} \end{cases}$$

Each interval, CI_0 or CI_1, has been proved that if it were used unconditionally for all datasets (i.e. without first checking $\widehat{\beta}_1$), it would capture β_0 with probability $1-\alpha$:

$$P\big(\beta_0 \in \mathrm{CI}_0\big) \;=\; 1-\alpha \quad \text{and} \quad P\big(\beta_0 \in \mathrm{CI}_1\big) \;=\; 1-\alpha,$$

However, once we introduce the random event T (i.e. "$|\widehat{\beta}_1| < c$"), the procedure conditions on whether T or T^c occurred. Hence, the overall coverage probability of β_0 is

$$\begin{aligned} P\Big(\beta_0 \in \mathrm{CI}_{\text{selected}}\Big) &= P\big(\beta_0 \in \mathrm{CI}_0 \;\wedge\; T\big) \;+\; P\big(\beta_0 \in \mathrm{CI}_1 \;\wedge\; T^c\big) \\ &= P(T)\,P\Big(\beta_0 \in \mathrm{CI}_0 \;\Big|\; T\Big) \;+\; P(T^c)\,P\Big(\beta_0 \in \mathrm{CI}_1 \;\Big|\; T^c\Big). \end{aligned}$$

Here, CI_0 and CI_1 were each derived under an assumption of unconditional use. In Method (c), however, each interval is used only after a data-driven event has occurred (T or T^c). Because $\widehat{\beta}_1$ is correlated with $\widehat{\beta}_0$ (and thus with the coverage event), the probabilities $P\big(\beta_0 \in \mathrm{CI}_0 \mid T\big)$ and $P\big(\beta_0 \in \mathrm{CI}_1 \mid T^c\big)$ are not guaranteed to be $1-\alpha$. This leads to post-model-selection inference. In most cases, the actual coverage will fall below 95%, unless adjustments.

For Method (d)

The key difference in Method (d) is that the data used to decide whether $\beta_1 = 0$ (the "test set") is disjoint from the data used to construct the confidence interval for β_0 (the "estimation set").

Let $(x_i, y_i)_{i=1}^{n_1}$ be the test set and $(x_i, y_i)_{i=n_1+1}^{n_1+n_2}$ be the estimation set, with $n_1 + n_2 = n$. We define:

$$T \;=\; \big\{|\hat{\beta}_1^{(\text{test})}| < c\big\},$$

where $\hat{\beta}_1^{(\text{test})}$ is the slope estimate using only the test set. Then we choose either

$$\mathrm{CI}_0(\text{est. data}) \quad \text{or} \quad \mathrm{CI}_1(\text{est. data}),$$

depending on whether T occurs.

Because the estimation set is independent of the test set, the distribution of the OLS estimates $\hat{\beta}_0^{(\text{est})}$ or $(\hat{\beta}_0^{(\text{est})}, \hat{\beta}_1^{(\text{est})})$ is unaffected by the event T. Hence if each CI (constant-model or linear-model) is constructed at nominal $(1-\alpha)$-level *conditionally* on the estimation data alone, it remains valid even when restricted to the event T or T^c. Formally,

In the data-splitting approach, we partition the full sample into:

$$\underbrace{\{(x_i, y_i)\}_{i=1}^{n_1}}_{\text{Test Set}} \quad \text{and} \quad \underbrace{\{(x_i, y_i)\}_{i=n_1+1}^{n_1+n_2}}_{\text{Estimation Set}},$$

with $n_1 + n_2 = n$. Let us define the event

$$T \;=\; \{|\hat{\beta}_1^{(\text{test})}| < c\},$$

where $\hat{\beta}_1^{(\text{test})}$ is the slope estimate obtained *only* from the test set. Once we observe whether T occurs, we decide which model (constant or linear) to use on the estimation set, and then construct the corresponding confidence interval for β_0.

Therefore, the random variable $\hat{\beta}_1^{(\text{test})}$ depends solely on the test-set observations $\{(x_i, y_i)\}_{i=1}^{n_1}$. The estimator $\hat{\beta}_0^{(\text{est})}$ depends solely on the disjoint estimation-set observations $\{(x_i, y_i)\}_{i=n_1+1}^{n}$. By design, these two subsets are *independent* samples from the same underlying distribution (e.g., i.i.d. draws). Consequently,

$$\hat{\beta}_1^{(\text{test})} \;\perp\; \hat{\beta}_0^{(\text{est})},$$

and so the event $T = \{|\hat{\beta}_1^{(\text{test})}| < c\}$ is also independent of $\hat{\beta}_0^{(\text{est})}$. Formally, for any Borel set A,

$$P\big(\hat{\beta}_0^{(\text{est})} \in A \;\big|\; T\big) \;=\; P\big(\hat{\beta}_0^{(\text{est})} \in A\big).$$

Now, suppose we define a $(1-\alpha)$-level confidence interval for β_0 using only the estimation set, under each model:

$$\mathrm{CI}_0^{(\text{est})} \quad \text{and} \quad \mathrm{CI}_1^{(\text{est})}.$$

Because $\mathrm{CI}_0^{(\text{est})}$ (resp. $\mathrm{CI}_1^{(\text{est})}$) is derived under the assumption that $\beta_1 = 0$ (resp. $\beta_1 \neq 0$) but only uses the estimation set, it has nominal coverage $1-\alpha$ *unconditionally* on that set. Due to independence from the test set, this property also holds conditionally on T. As we have shown in part(a) and (b):

$$P\Big(\beta_0 \in \mathrm{CI}_0^{(\text{est})} \;\Big|\; T\Big) = P\Big(\beta_0 \in \mathrm{CI}_0^{(\text{est})}\Big) \;=\; 1-\alpha,$$

and similarly for $\mathrm{CI}_1^{(\text{est})}$ given T^c. Therefore, the marginal coverage rate in Method (d) is approximately 95%.

Hence,

$$\begin{aligned} P\big(\beta_0 \in \mathrm{CI}_{\text{selected}}\big) &= P(T)\,P\big(\beta_0 \in \mathrm{CI}_0^{(\text{est})} \mid T\big) \;+\; P(T^c)\,P\big(\beta_0 \in \mathrm{CI}_1^{(\text{est})} \mid T^c\big) \\ &= P(T)\,(1-\alpha) \;+\; P(T^c)\,(1-\alpha) \;=\; 1-\alpha. \end{aligned}$$

Thus, by splitting the data, the model selection event is separated from the interval construction, preserving the nominal coverage level $(1-\alpha)$ for β_0. This explains why Method (d) achieves overall coverage or marginal coverage closer to 95% compared to Method (c).