HW0324

Yung-Jung Cheng

2025-03-30

Q1

Let $K = 2$, show that (b_1, b_2) in p.29 of slide in Ch 5 reduces to the formual of (b_1, b_2) in (2.7)-(2.8)

Ans

Step 1: Model and Objective Function

Consider the simple linear regression model:

$$y_i = b_1 + b_2 x_i + e_i, \quad i = 1, 2, \ldots, N$$

The least squares method minimizes the sum of squared errors (SSE):

$$SSE(b_1, b_2) = \sum_{i=1}^{N} (y_i - b_1 - b_2 x_i)^2$$

Step 2: Take Derivatives with Respect to b_1 and b_2

Partial derivative w.r.t. b_1:

$$\frac{\partial SSE}{\partial b_1} = -2 \sum (y_i - b_1 - b_2 x_i) = 0$$

Solving:

$$\sum y_i = N b_1 + b_2 \sum x_i \quad \Rightarrow \quad b_1 = \overline{y} - b_2 \overline{x}$$

Partial derivative w.r.t. b_2:

$$\frac{\partial SSE}{\partial b_2} = -2 \sum x_i (y_i - b_1 - b_2 x_i) = 0$$

Substitute $b_1 = \overline{y} - b_2 \overline{x}$:

$$\sum x_i (y_i - \overline{y} + b_2 \overline{x} - b_2 x_i) = 0$$

Simplifying:

$$\sum (x_i - \overline{x})(y_i - \overline{y}) = b_2 \sum (x_i - \overline{x})^2$$

Thus:

$$b_2 = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2}$$

Conclusion from LSE Derivation:

We obtain the closed-form solution:

$$b_2 = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2}, \quad b_1 = \overline{y} - b_2\overline{x}$$

Verify Equivalence with Matrix Formula $b = (X'X)^{-1}X'Y$

Define:

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix}, \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

Compute:

$$X'X = \begin{bmatrix} N & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}, \quad X'Y = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

Then:

$$(X'X)^{-1} = \frac{1}{N\sum x_i^2 - (\sum x_i)^2}\begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & N \end{bmatrix}$$

So:

$$b = (X'X)^{-1}X'Y = \frac{1}{D}\begin{bmatrix} \sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i \\ N\sum x_i y_i - \sum x_i \sum y_i \end{bmatrix}$$

Where $D = N\sum x_i^2 - (\sum x_i)^2$.

Final Transformation

Using the identity:

$$\sum(x_i - \overline{x})(y_i - \overline{y}) = \sum x_i y_i - N\overline{x}\,\overline{y}$$

We can show:

$$b_2 = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2}, \quad b_1 = \overline{y} - b_2\overline{x}$$

Conclusion

We have shown that the scalar-form least squares solution (b_1, b_2) derived from minimizing SSE is equivalent to the matrix-form solution:

$$b = (X'X)^{-1}X'Y$$

for the case $K = 2$

Q2

Let $K = 2$, show that $cov(b_1, b_2)$ in p.30 of slide in Ch 5 reduces to the formula of in (2.14)-(2.16)

Ans

Step 1: General variance formula for OLS estimators

Given the linear model:

$$Y = X\beta + \varepsilon, \quad \text{with } \mathrm{Var}(\varepsilon) = \sigma^2 I,$$

the least squares estimator is:

$$\hat{\beta} = (X'X)^{-1}X'Y$$

and its variance-covariance matrix is:

$$\mathrm{Var}(\hat{\beta}) = \sigma^2 (X'X)^{-1}$$

Step 2: Compute $(X'X)^{-1}$ for simple linear regression ($K = 2$)

Let:

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix}$$

Then:

$$X'X = \begin{bmatrix} N & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}$$

Using the 2×2 inverse formula:

$$(X'X)^{-1} = \frac{1}{D} \begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & N \end{bmatrix}, \quad \text{where } D = N \sum x_i^2 - (\sum x_i)^2$$

Step 3: Multiply by σ^2 to get the covariance matrix

$$\mathrm{Var}(b) = \sigma^2 (X'X)^{-1} = \frac{\sigma^2}{D} \begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & N \end{bmatrix}$$

Extract components:

- $\mathrm{Var}(b_1) = \frac{\sigma^2 \sum x_i^2}{D}$
- $\mathrm{Var}(b_2) = \frac{\sigma^2 N}{D}$
- $\mathrm{Cov}(b_1, b_2) = -\frac{\sigma^2 \sum x_i}{D}$

Step 4: Express D using centered data

Define:

$$\overline{x} = \frac{1}{N}\sum x_i, \quad S_{xx} = \sum (x_i - \overline{x})^2$$

Then:

$$D = N\sum x_i^2 - (\sum x_i)^2 = NS_{xx}$$

Step 5: Express variances in terms of S_{xx} and $\overline{x}$

Use identity:

$$\sum x_i^2 = S_{xx} + N\overline{x}^2$$

Substitute into each term:

- $\operatorname{Var}(b_1) = \frac{\sigma^2(S_{xx}+N\overline{x}^2)}{NS_{xx}} = \frac{\sigma^2}{N} + \frac{\sigma^2\overline{x}^2}{S_{xx}}$
- $\operatorname{Var}(b_2) = \frac{\sigma^2}{S_{xx}}$
- $\operatorname{Cov}(b_1, b_2) = -\frac{\sigma^2\overline{x}}{S_{xx}}$

Final Result (as in page 30)

$$\operatorname{Var}(b_1) = \frac{\sigma^2}{N} + \frac{\sigma^2\overline{x}^2}{\sum(x_i - \overline{x})^2}$$

$$\operatorname{Var}(b_2) = \frac{\sigma^2}{\sum(x_i - \overline{x})^2}$$

$$\operatorname{Cov}(b_1, b_2) = -\frac{\sigma^2\overline{x}}{\sum(x_i - \overline{x})^2}$$

These match exactly with the expressions on slide page 30.

Q03

Consider the following model that relates the percentage of a household's budget spent on alcohol $WALC$ to total expenditure $TOTEXP$, age of the household head AGE, and the number of children in the household NK:

$$WALC = \beta_1 + \beta_2 \ln(TOTEXP) + \beta_3 NK + \beta_4 AGE + e$$

This model was estimated using 1200 observations from London. An incomplete version of this output is provided in Table 5.6.

TABLE 5.6 Output for Exercise 5.3

Dependent Variable: WALC
Included observations: 1200

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.4515	2.2019	?	0.5099
ln(TOTEXP)	2.7648	?	5.7103	0.0000
NK	?	0.3695	-3.9376	0.0001
AGE	-0.1503	0.0235	-6.4019	0.0000

- R-squared: ?
- S.E. of regression: 6.39547
- Sum squared resid: 46221.62
- Mean dependent var: 6.19434
- S.D. dependent var: 6.39547

Q3(a)

Fill in the following blank spaces that appear in this table.

i. The t-statistic for β_1
ii. The standard error for β_2
iii. The estimate β_3
iv. R^2
v. $\hat{\sigma}$

Ans

i. The t-statistic for β_1 $\boxed{0.6593}$

$$t = \frac{\hat{\beta}_1}{\text{SE}(\hat{\beta}_1)} = \frac{1.4515}{2.2019} \approx 0.6593$$

ii. The standard error for β_2 $\boxed{0.4840}$

$$\text{SE} = \frac{\hat{\beta}_2}{t} = \frac{2.7648}{5.7103} \approx 0.4840$$

iii. The estimate β_3 $\boxed{-1.4548}$

$$\hat{\beta}_3 = t \cdot \text{SE} = (-3.9376)(0.3695) \approx -1.4548$$

iv. R^2 $\boxed{0.0588}$

$$R^2 = 1 - \frac{SSR}{TSS}$$

- SSR = 46221.62
- TSS = $n \cdot \text{S.D.}^2 = 1200 \cdot (6.39547)^2 \approx 49107.84$

$$R^2 = 1 - \frac{46221.62}{49107.84} \approx 1 - 0.9412 = 0.0588$$

v. $\hat{\sigma}$ $\boxed{6.2180}$

$$\hat{\sigma} = \sqrt{\frac{SSR}{n-K-1}} = \sqrt{\frac{46221.62}{1196}} = \sqrt{38.6529} \approx 6.2180$$

Q3(b)

Interpret each of the estimates β_2, β_3, and β_4.

Ans

Variable	Marginal Effect Formula	Economic Interpretation
$\hat{\beta}_2$	$\partial WALC/\partial \ln(TOTEXP)$	A 1% increase in total expenditure increases alcohol share by 2.7648 percentage points.
$\hat{\beta}_3$	$\partial WALC/\partial NK$	One additional child reduces alcohol share by 1.4548 percentage points.
$\hat{\beta}_4$	$\partial WALC/\partial AGE$	One more year of age reduces alcohol share by 0.1503 percentage points.

Q3(c)

Compute a 95% interval estimate for β_4. What does this interval tell you?

Ans

We are given: - $\hat{\beta}_4 = -0.1503$ - Standard error $SE(\hat{\beta}_4) = 0.0235$ - $z_{0.025} = 1.96$ for a 95% confidence level

Step 1: Compute the margin of error
$ME = 1.96 \times 0.0235 = 0.0461$

Step 2: Construct the confidence interval
Lower bound: $-0.1503 - 0.0461 = -0.1964$
Upper bound: $-0.1503 + 0.0461 = -0.1042$

Conclusion:
The 95% confidence interval for β_4 is $(-0.1964, -0.1042)$.
We are 95% confident that the effect of age on alcohol expenditure share lies within this interval.
Since the entire interval is negative, this implies that age has a statistically significant negative effect.

Q3(d)

Are each of the coefficient estimates significant at a 5% level? Why?

Ans

To determine whether each coefficient is statistically significant at the 5% level, we compare their p-values with 0.05.

Coefficient	Estimate	t-Statistic	p-value	Significant at 5%?
β_1	1.4515	0.6593	0.5099	No
β_2	2.7648	5.7103	0.0000	Yes
β_3	-1.4548	-3.9376	0.0001	Yes
β_4	-0.1503	-6.4019	0.0000	Yes

Conclusion:
At the 5% significance level, β_2, β_3, and β_4 are statistically significant, while β_1 is not.

Q3(e)

Test the hypothesis that the addition of an extra child decreases the mean budget share of alcohol by 2 percentage points against the alternative that the decrease is not equal to 2 percentage points. Use a 5% significance level.

Ans

Null hypothesis: $H_0 : \beta_3 = -2$
Alternative hypothesis: $H_1 : \beta_3 \neq -2$
Significance level: $\alpha = 0.05$

We are given: - $\hat{\beta}_3 = -1.4548$ - $SE(\hat{\beta}_3) = 0.3695$

Step 1: Compute the test statistic
$t = \frac{-1.4548-(-2)}{0.3695} = \frac{0.5452}{0.3695} \approx 1.4755$

Step 2: Compare with critical value
The critical value at 5% significance level (two-tailed) is 1.96.
Since $|t| = 1.4755 < 1.96$, we fail to reject the null hypothesis.

Conclusion:
There is not enough evidence at the 5% level to conclude that β_3 differs from -2.

Q23

The file `cocaine` contains 56 observations on variables related to sales of cocaine powder in northeastern California over the period 1984–1991. The data are a subset of those used in the study: Caulkins, J. P. and R. Padman (1993), "Quantity Discounts and Quality Premia for Illicit Drugs," *Journal of the American Statistical Association*, 88, 748–757. The variables are: - `PRICE` : price per gram in dollars for a cocaine sale
- `QUANT` : number of grams of cocaine in a given sale
- `QUAL` : quality of the cocaine expressed as percentage purity
- `TREND` : a time variable with 1984 = 1 up to 1991 = 8
Consider the regression model:

$$PRICE = \beta_1 + \beta_2 QUANT + \beta_3 QUAL + \beta_4 TREND + e$$

Q23(a)

What signs would you expect on the coefficients β_2, β_3, and β_4?

Ans

- β_2: Expected to be **negative** due to quantity discounts — as quantity increases, price per gram decreases.
- β_3: Expected to be **positive** because higher purity should command a higher price.
- β_4: Likely **negative** if price has declined over time due to increased availability or law enforcement pressure.

Q23(b)

Use your computer software to estimate the equation. Report the results and interpret the coefficient estimates. Have the signs turned out as you expected?

Ans

```
model <- lm(price ~ quant + qual + trend, data = cocaine)
summary(model)
```

```
## 
## Call:
## lm(formula = price ~ quant + qual + trend, data = cocaine)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -43.479 -12.014  -3.743  13.969  43.753 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 90.84669    8.58025  10.588 1.39e-14 ***
## quant       -0.05997    0.01018  -5.892 2.85e-07 ***
## qual         0.11621    0.20326   0.572   0.5700    
## trend       -2.35458    1.38612  -1.699   0.0954 .  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 20.06 on 52 degrees of freedom
## Multiple R-squared:  0.5097, Adjusted R-squared:  0.4814 
## F-statistic: 18.02 on 3 and 52 DF,  p-value: 3.806e-08
```

The estimated regression equation is:

$$\widehat{PRICE} = 90.8467 - 0.0600 \cdot QUANT + 0.1162 \cdot QUAL - 2.3546 \cdot TREND$$

Interpretation:

- **QUANT**: The coefficient is -0.0600 and is statistically significant at the 1% level. This confirms the presence of quantity discounts: as quantity increases by one gram, the price per gram drops by 0.06 dollars.
- **QUAL**: The coefficient is 0.1162, which has the expected positive sign, suggesting higher purity leads to higher price. However, it is **not statistically significant** (p = 0.5700), so we cannot confidently say that purity affects price.
- **TREND**: The coefficient is -2.3546 with a p-value of 0.0954, which is marginally significant at the 10% level. This implies a possible downward trend in price over time.

Conclusion:
The signs of all coefficients match the expectations from part (a), but only **QUANT** is strongly significant. **QUAL** is not significant, and **TREND** is weakly significant at the 10% level.

Q23(c)

What proportion of variation in cocaine price is explained jointly by variation in quantity, quality, and time?

Ans

The model's R-squared is 0.5097, meaning that approximately **50.97%** of the variation in cocaine prices is explained by the variation in quantity, quality, and time (trend).

Interpretation:
This suggests that about half of the price variation can be accounted for by these three variables. The remaining variation may be due to other factors not included in the model, such as transaction location, seller identity, buyer type, or law enforcement pressure.

Q23(d)

It is claimed that the greater the number of sales, the higher the risk of getting caught. Thus, sellers are willing to accept a lower price if they can make sales in larger quantities.
Set up H_0 and H_1 that would be appropriate to test this hypothesis. Carry out the hypothesis test.

Ans

Claim: Sellers are willing to accept a lower price when selling in larger quantities.

We test:

- Null hypothesis: $H_0 : \beta_2 = 0$
- Alternative hypothesis: $H_1 : \beta_2 < 0$
- Significance level: $\alpha = 0.05$

From the regression output: - Estimate: $\hat{\beta}_2 = -0.05997$ - Standard error: $SE = 0.01018$ - Test statistic: $t = -5.892$ - p-value: 2.85×10^{-7}

Conclusion:
Since the p-value is far below 0.05 and the t-statistic is less than the critical value, we reject the null hypothesis.
There is strong evidence to support the existence of a quantity discount in cocaine pricing.

Q23(e)

Test the hypothesis that the quality of cocaine has no influence on expected price against the alternative that a premium is paid for better-quality cocaine.

Ans

We test whether better-quality cocaine commands a higher price.

- Null hypothesis: $H_0 : \beta_3 = 0$
- Alternative hypothesis: $H_1 : \beta_3 > 0$
- Significance level: $\alpha = 0.05$

From the regression output: - Estimate: $\hat{\beta}_3 = 0.1162$ - Standard error: $SE = 0.2033$ - Test statistic: $t = 0.572$ - Critical value: $t_{0.05,52} \approx 1.675$ - p-value = 0.5700

Conclusion:
We fail to reject the null hypothesis.
There is no statistically significant evidence that a premium is paid for better-quality cocaine.

Q23(f)

What is the average annual change in the cocaine price? Can you suggest why price might be changing in this direction?

Ans

From the regression results, the coefficient on the TREND variable is:

$$\hat{\beta}_4 = -2.3546$$

This indicates that the average cocaine price per gram **decreased by approximately $2.35 per year** from 1984 to 1991.

Possible Reasons: - Increased supply due to mass production and export from South America (e.g., Colombia) - Competition among dealers driving prices down - Law enforcement unable to constrain supply effectively - Market saturation and commoditization of cocaine - Economic or social factors reducing willingness to pay

Conclusion:
There was a significant downward trend in cocaine prices over this period, likely driven by supply-side factors and increased market competition.