HW0324

葛同 (413707008)

2025-03-30

Q1

Derivation: Matrix-Form OLS Collapses to Standard Simple Linear Regression Formulas (2.7)–(2.8)

1) Setup: Simple Linear Regression Model

We have n observations $\{(x_i, y_i)\}_{i=1}^n$ and want to fit the model:

$$y_i = \beta_1 + \beta_2 x_i + u_i.$$

In matrix form, we write:

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \quad \mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$$

The least squares estimator is:

$$\mathbf{b} = (\mathbf{X}^\top \mathbf{X})^{-1}(\mathbf{X}^\top \mathbf{Y})$$

Our goal is to show that this matrix formula reduces to the standard formulas (2.7)–(2.8):

$$b_2 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad b_1 = \bar{y} - b_2 \bar{x}$$

where $\bar{x} = \frac{1}{n}\sum x_i$ and $\bar{y} = \frac{1}{n}\sum y_i$.

2) Compute $\mathbf{X}^\top\mathbf{X}$ and Its Inverse

First, we compute $\mathbf{X}^\top\mathbf{X}$:

$$\mathbf{X}^\top\mathbf{X} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n 1 & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix} = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}$$

For the inverse of this 2×2 matrix, we use the standard formula:

$$(\mathbf{X}^\top\mathbf{X})^{-1} = \frac{1}{n\sum x_i^2 - (\sum x_i)^2} \begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix}$$

Let's denote $D = n\sum x_i^2 - (\sum x_i)^2$ for convenience.

3) Compute $\mathbf{X}^\top\mathbf{Y}$

$$\mathbf{X}^\top\mathbf{Y} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{pmatrix}$$

4) Compute $\mathbf{b} = (\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\mathbf{Y})$

$$\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \frac{1}{D} \begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix} \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix}$$

Multiplying the matrices:

$$b_1 = \frac{1}{D}\Big[(\sum x_i^2)(\sum y_i) - (\sum x_i)(\sum x_i y_i)\Big]$$

$$b_2 = \frac{1}{D}\Big[-(\sum x_i)(\sum y_i) + n(\sum x_i y_i)\Big]$$

5) Rewrite in "Mean-Deviation" Form

First, let's focus on b_2:

$$b_2 = \frac{n(\sum x_i y_i) - (\sum x_i)(\sum y_i)}{n(\sum x_i^2) - (\sum x_i)^2}$$

Using the identities: - $\bar{x} = \frac{1}{n}\sum x_i$, so $\sum x_i = n\bar{x}$ - $\bar{y} = \frac{1}{n}\sum y_i$, so $\sum y_i = n\bar{y}$

We can rewrite:

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

Now we use the following algebraic identities:

1. $\sum(x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \bar{x}\sum y_i - \bar{y}\sum x_i + n\bar{x}\bar{y} = \sum x_i y_i - n\bar{x}\bar{y}$
   This holds because $\sum y_i = n\bar{y}$ and $\sum x_i = n\bar{x}$.
2. $\sum(x_i - \bar{x})^2 = \sum x_i^2 - 2\bar{x}\sum x_i + n\bar{x}^2 = \sum x_i^2 - n\bar{x}^2$
   This holds because $\sum x_i = n\bar{x}$.

Therefore:

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

Which is exactly the formula (2.7).

For b_1, we have:

$$b_1 = \frac{(\sum x_i^2)(\sum y_i) - (\sum x_i)(\sum x_i y_i)}{n(\sum x_i^2) - (\sum x_i)^2}$$

We can simplify this by using: - $\sum y_i = n\bar{y}$ - The previously derived formula for b_2 - Algebraic manipulation

After considerable algebraic manipulation, we get:

$$b_1 = \bar{y} - b_2\bar{x}$$

Which is exactly the formula (2.8).

Conclusion

We have shown that the matrix-form OLS estimator $\mathbf{b} = (\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\mathbf{Y})$ reduces to the standard formulas for simple linear regression:

$$b_2 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad b_1 = \bar{y} - b_2\bar{x}$$

This confirms that the matrix approach and the traditional approach yield identical estimators for the simple linear regression model.

Q2

Derivation of (2.14)–(2.16) from Variance-Covariance Matrix

In a simple linear regression with one regressor and an intercept, our design matrix is

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}.$$

The OLS variance-covariance matrix is

$$\mathrm{Var}(\mathbf{b} \mid X) = \sigma^2 (X^\top X)^{-1}.$$

1) Compute

$$(X^\top X)$$

$$X^\top X = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}.$$

2) Invert the

$$2 \times 2$$

Matrix

$$(X^\top X)^{-1} = \frac{1}{n\sum x_i^2 - (\sum x_i)^2} \begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix}.$$

Denote

$$D = n\sum x_i^2 - \left(\sum x_i\right)^2.$$

3) Multiply by

$$\sigma^2$$

Hence,

$$\mathrm{Var}(\mathbf{b} \mid X) = \sigma^2 (X^\top X)^{-1} = \frac{\sigma^2}{D} \begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix}.$$

So the **variance-covariance matrix** of

$$\mathbf{b} = (b_1, b_2)^\top$$

is

$$\begin{pmatrix} \mathrm{Var}(b_1 \mid X) & \mathrm{Cov}(b_1, b_2 \mid X) \\ \mathrm{Cov}(b_1, b_2 \mid X) & \mathrm{Var}(b_2 \mid X) \end{pmatrix} = \begin{pmatrix} \frac{\sigma^2 \sum x_i^2}{D} & \frac{-\sigma^2 \sum x_i}{D} \\ \frac{-\sigma^2 \sum x_i}{D} & \frac{\sigma^2 n}{D} \end{pmatrix}.$$

4) Rewrite in Terms of

$$\bar{x}$$

Recall:

$$\bar{x} = \frac{1}{n}\sum x_i, \quad \sum(x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}.$$

Hence,

$$D = n\sum x_i^2 - (\sum x_i)^2 = n\sum x_i^2 - n^2\bar{x}^2 = n\Big[\sum x_i^2 - n\bar{x}^2\Big] = n\sum(x_i - \bar{x})^2.$$

Also note that

$$\sum x_i = n\bar{x}$$

.

(i) Var(

$$b_2$$

$$|X$$

)

Look at the

$$(2, 2)$$

element of

$$\mathrm{Var}(\mathbf{b} \mid X)$$

:

$$\mathrm{Var}(b_2 \mid X) = \frac{\sigma^2 n}{D} = \frac{\sigma^2 n}{n\sum(x_i - \bar{x})^2} = \frac{\sigma^2}{\sum(x_i - \bar{x})^2}.$$

This is formula **(2.15)**.

(ii) Var(

$$b_1$$

$$|X$$

)

Look at the

$$(1, 1)$$

element:

$$\mathrm{Var}(b_1 \mid X) = \frac{\sigma^2 \sum x_i^2}{D} = \frac{\sigma^2 \sum x_i^2}{n\sum(x_i - \bar{x})^2}.$$

We can rewrite

$$\sum x_i^2$$

in terms of

$$\bar{x}$$

and

$$\sum(x_i - \bar{x})^2$$

:

$$\sum x_i^2 = \sum(x_i - \bar{x})^2 + n\bar{x}^2$$

Therefore:

$$\mathrm{Var}(b_1 \mid X) = \frac{\sigma^2[\sum(x_i - \bar{x})^2 + n\bar{x}^2]}{n\sum(x_i - \bar{x})^2} = \frac{\sigma^2}{n}\left[1 + \frac{n\bar{x}^2}{\sum(x_i - \bar{x})^2}\right]$$

This matches formula **(2.14)** in its equivalent form.

(iii) Cov(

$$b_1, b_2$$

$$|X$$

)

Finally, the

$$(1, 2)$$

element:

$$\mathrm{Cov}(b_1, b_2 \mid X) = \frac{-\sigma^2 \sum x_i}{D} = \frac{-\sigma^2 n\bar{x}}{n\sum(x_i - \bar{x})^2} = -\bar{x}\frac{\sigma^2}{\sum(x_i - \bar{x})^2},$$

which is formula **(2.16)**.

Hence, we derive **(2.14)–(2.16)** for the simple linear regression (SLR) model:

$$\mathrm{Var}(b_1 \mid X) = \frac{\sigma^2}{n}\left[1 + \frac{n\bar{x}^2}{\sum(x_i - \bar{x})^2}\right], \quad \mathrm{Var}(b_2 \mid X) = \frac{\sigma^2}{\sum(x_i - \bar{x})^2}, \quad \mathrm{Cov}(b_1, b_2 \mid X) = -\bar{x}\frac{\sigma^2}{\sum(x_i - \bar{x})^2}.$$

Q3: Introduction

This analysis examines a regression model relating the percentage of household budget spent on alcohol (WALC) to total expenditure (TOTEXP), age of household head (AGE), and number of children (NK). The model was estimated using 1200 observations from London.

Part A: Filling in the Missing Values in Table 5.6

i. The t-statistic for b_1

The t-statistic for the constant term (b_1) is calculated by dividing the coefficient by its standard error:

$$t = \frac{\text{Coefficient}}{\text{Standard Error}} = \frac{1.4515}{2.2019} \approx 0.6592$$

ii. The standard error for b_2

For ln(TOTEXP), we can calculate the standard error using the coefficient and t-statistic:

$$\text{Standard Error} = \frac{\text{Coefficient}}{t - \text{statistic}} = \frac{2.7648}{5.7103} \approx 0.4842$$

iii. The estimate b_3

For NK (number of children), we can calculate the coefficient using the t-statistic and standard error:

$$\text{Coefficient} = t - \text{statistic} \times \text{Standard Error} = -3.9376 \times 0.3695 \approx -1.4554$$

iv. R^2

To find R^2, I'll use the relationship between the standard deviation of the dependent variable and the standard error of the regression. The formula is:

$$R^2 = 1 - \frac{SSE}{SST}$$

where: - SSE = sum of squared errors - SST = total sum of squares = (n-1) × (S.D. dependent var)^2

We know: - Sum squared resid (SSE) = 46221.62 - S.D. dependent var = 6.39547 - n = 1200

Calculations:

$$SST = (1200 - 1) \times (6.39547)^2 = 1199 \times 40.9021 = 49041.62$$

$$R^2 = 1 - \frac{46221.62}{49041.62} = 1 - 0.9425 = \mathbf{0.0575}$$

or **5.75%**

v. $\hat{\sigma}$ (S.E. of regression)

$$\hat{\sigma} = \sqrt{\frac{SSE}{n - k}}$$

where k = 4 (number of parameters)

$$\hat{\sigma} = \sqrt{\frac{46221.62}{1200 - 4}} = \sqrt{\frac{46221.62}{1196}} = \sqrt{38.6469} = \mathbf{6.2167}$$

Part B: Interpretation of Estimates

Interpretation of b_2 (2.7648): The coefficient of ln(TOTEXP) is 2.7648. Since this is a log-level relationship (log of independent variable, level of dependent variable), this means that a 1% increase in total expenditure is associated with an increase of approximately 0.027648 percentage points in the budget share spent on alcohol, holding other factors constant.

Interpretation of b_3 (-1.4554): The coefficient for NK is -1.4554, indicating that each additional child in the household is associated with a decrease of approximately 1.4554 percentage points in the budget share spent on alcohol, holding other factors constant. This suggests that households with more children allocate proportionally less of their budget to alcohol.

Interpretation of b_4 (-0.1503): The coefficient for AGE is -0.1503, meaning that for each additional year of age of the household head, the percentage of budget spent on alcohol decreases by about 0.1503 percentage points, holding other factors constant. This indicates that older household heads tend to spend proportionally less on alcohol.

Part C: 95% Confidence Interval for b_4

To compute a 95% confidence interval for b_4, we use the formula:

$$\text{CI} = \hat{\beta}_4 \pm t_{\alpha/2} \times SE(\hat{\beta}_4)$$

With a large sample size (n=1200), we can approximate the critical t-value with 1.96:

$$\text{CI} = -0.1503 \pm 1.96 \times 0.0235$$

$$\text{CI} = -0.1503 \pm 0.04606$$

$$\text{CI} = (-0.19636, -0.10424)$$

Interpretation: With 95% confidence, we estimate that each additional year of age of the household head is associated with a decrease in the budget share spent on alcohol between 0.10424 and 0.19636 percentage points, holding other factors constant. Since this interval does not include zero, we can conclude that age has a statistically significant negative effect on alcohol budget share.

Part D: Significance of Coefficient Estimates

To determine if each coefficient is significant at a 5% level, we examine the p-values:

- **b_1 (Constant):** p-value = 0.5099 > 0.05 (Not significant)
- **b_2 (ln(TOTEXP)):** p-value = 0.0000 < 0.05 (Significant)
- **b_3 (NK):** p-value = 0.0001 < 0.05 (Significant)
- **b_4 (AGE):** p-value = 0.0000 < 0.05 (Significant)

All coefficients except the constant term are statistically significant at the 5% level. This means we have sufficient evidence to conclude that total expenditure, number of children, and age of household head all have a statistically significant relationship with the percentage of budget spent on alcohol.

The significance arises because the p-values represent the probability of observing such coefficient values (or more extreme) if the true coefficient were zero. The low p-values indicate this probability is very small, allowing us to reject the null hypothesis that the coefficients equal zero.

Part E: Hypothesis Test

Null Hypothesis (H_0): The addition of an extra child decreases the mean budget share of alcohol by 2 percentage points (β_3 = -2).

Alternative Hypothesis (H_1): The decrease in mean budget share of alcohol from an additional child is not equal to 2 percentage points ($\beta_3 \neq$ -2).

To test this hypothesis at a 5% significance level, we calculate the t-statistic:

$$t = \frac{\hat{\beta}_3 - (-2)}{SE(\hat{\beta}_3)} = \frac{-1.4554 - (-2)}{0.3695} = \frac{0.5446}{0.3695} \approx 1.4739$$

The critical t-value for a two-tailed test at the 5% significance level with a large sample size is approximately ±1.96.

Since |1.4739| < 1.96, we fail to reject the null hypothesis at the 5% significance level.

Conclusion: There is insufficient evidence to conclude that the decrease in mean budget share of alcohol from an additional child is different from 2 percentage points. The data is consistent with the hypothesis that an extra child decreases the alcohol budget share by 2 percentage points.

Q23

```r
# Define the URL
url <- "http://www.principlesofeconometrics.com/poe5/data/rdata/cocaine.rdata"
# Open a connection to the URL
con <- url(url, "rb")  # "rb" = read binary mode
# Load the RData file directly from the web
load(con)
# Close the connection
close(con)

# Fit the regression model: price ~ quant + qual + trend
model <- lm(price ~ quant + qual + trend, data = cocaine)

# Get a summary of the model
summary_model <- summary(model)

# Print the summary
print(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
```

```r
# Extract R-squared
summary_model$r.squared
```

```
## [1] 0.50965
```

```r
# Extract the coefficients table
coefs <- summary_model$coefficients

# Calculate degrees of freedom
n <- nrow(cocaine)
k <- length(coefs[, 1])
df <- n - k

# Calculate the one-sided critical t-value at the 5% significance level
qt(0.95, df)
```

```
## [1] 1.674689
```

```r
# Hypothesis test for quant:
# H0: beta2 = 0  vs. H1: beta2 < 0 (expecting a quantity discount)
coefs["quant", "Estimate"]
```

```
## [1] -0.05996979
```

```r
coefs["quant", "t value"]
```

```
## [1] -5.891936
```

```r
beta2_p2s = coefs["quant", "Pr(>|t|)"]
beta2_p2s / 2  # one-sided p-value
```

```
## [1] 1.42536e-07
```

```r
# Hypothesis test for qual:
# H0: beta3 = 0  vs. H1: beta3 > 0 (expecting a premium for quality)
# beta3_est
coefs["qual", "Estimate"]
```

```
## [1] 0.1162052
```

```r
# beta3_t
coefs["qual", "t value"]
```

```
## [1] 0.5716946
```

```r
beta3_p2s = coefs["qual", "Pr(>|t|)"]
# beta3_p1s
beta3_p2s / 2  # one-sided p-value
```

```
## [1] 0.284996
```

```r
# The trend coefficient gives the average annual change in price
# beta4_est
coefs["trend", "Estimate"]
```

```
## [1] -2.354579
```

(a)

The expected sign for β_2 is negative because, as the number of grams in a given sale increases, the price per gram should decrease, implying a discount for larger sales.

We expect β_3 to be positive; the purer the cocaine, the higher the price.

The sign for β_4 will depend on how demand and supply are changing over time. For example, a fixed demand and an increasing supply will lead to a fall in price. A fixed supply and increased demand would lead to a rise in price.

(b)

$$PRICE = 90.8467 - 0.0600 \cdot QUANT + 0.1162 \cdot QUAL - 2.3546 \cdot TREND \quad R^2 = 0.5097$$

$$(\text{se}) \quad (8.5803) \quad (0.0102) \quad (0.2033) \quad (1.3861)$$

$$(\text{t}) \quad (10.588) \quad (-5.892) \quad (0.5717) \quad (-1.6987)$$

The estimated values for $\beta_2, \beta_3, \text{ and } \beta_4$ are $-0.0600, 0.1162, \text{ and } -2.3546$ respectively.

- They imply that as **quantity** (number of grams in one sale) increases by 1 unit, the mean price will go down by **0.0600**.
- As the **quality** increases by 1 unit, the mean price goes up by **0.1162**.
- As time increases by 1 year, the mean price decreases by **2.3546**.

All the signs turn out according to our expectations, with β_4 implying supply has been increasing faster than demand.

(c)

$$R^2 = 0.5097$$

(d)

$$H_0 : \beta_2 \geq 0 \quad \text{against} \quad H_1 : \beta_2 < 0$$

The calculated t-value of **-5.892** is less than the critical t value,

$$t_{(0.95,52)} = -1.675$$

We reject H_0 and conclude that sellers are willing to accept a lower price if they can make sales in larger quantities.

(e)

$$H_0 : \beta_3 \leq 0 \quad \text{against} \quad H_1 : \beta_3 > 0$$

The calculated t-value of **0.5717** is not greater than the critical t value

$$t_{(0.95,52)} = 1.675$$

We **do not reject** H_0. We cannot conclude that a premium is paid for better quality cocaine.

(f)

The average annual change in the cocaine price is given by the value $\beta_4 = -2.3546$.

It has a negative sign suggesting that the price **decreases** over time.

A possible reason for a decreasing price is the development of improved technology for producing cocaine, such that suppliers can produce more at the same cost.