HW0324 Pinyo – 312712017

HW0324Q1

Let show that (b1,b2) in p.29 of slide in Ch5 reduces to the formula of (b1,b2) in (2.7)-(2.8)

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

The Ordinary Least Squares (OLS) Estimators

$$b_2 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \tag{2.7}$$

$$b_1 = \bar{y} - b_2\bar{x} \tag{2.8}$$

where $\bar{y} = \sum y_i/N$ and $\bar{x} = \sum x_i/N$ are the sample means of the observations on y and x.

1. Matrix Representation of OLS

The OLS estimator for a linear regression model is given by:

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

where:

- X is the design matrix of independent variables (including a column of ones for the intercept),
- Y is the vector of dependent variable observations,
- b is the vector of estimated coefficients.

For a simple linear regression with one independent variable (x), X can be written as:

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix},$$

where b_1 is the intercept and b_2 is the slope.

**2. Expanding $X'X$ and $X'Y$

The transpose of X, denoted X', is:

$$X' = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix}.$$

Now compute:

Step 1: Compute $X'X$:

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

Step 2: Compute $X'Y$:

$$X'Y = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}.$$

3. Solving for $b = (X'X)^{-1}(X'Y)$

To find b, we need to compute $(X'X)^{-1}$. The inverse of $X'X$ is given by:

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

Now multiply $(X'X)^{-1}$ by $X'Y$:

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

Performing the matrix multiplication:

Intercept (b_1):

The first element of b, representing the intercept, becomes:

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

Slope (b_2):

The second element of b, representing the slope, becomes:

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

4. Simplifying to Match Formula (2.7 and 2.8)

Step 1: Simplify b_2:

The slope formula can be rewritten using deviations from the mean:

- Mean of y : $\bar{y} = \frac{\sum y_i}{n}$,
- Mean of x : $\bar{x} = \frac{\sum x_i}{n}$.

Using these means, rewrite b_2:

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

This matches equation (2.7).

Step 2: Derive b_1:

Once b_2 is known, substitute it into the formula for the intercept:

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

This matches equation (2.8).

HW0324Q2

Let show that cov(b1,b2) in p.30 of slide in Ch5 reduces to the formula in (2.14)-(2.16)

$$b \sim N(\beta, \sigma^2 (X'X)^{-1}),$$

$$\text{var}(b_1|\mathbf{x}) = \sigma^2 \left[\frac{\sum x_i^2}{N \sum (x_i - \bar{x})^2} \right] \tag{2.14}$$

$$\text{var}(b_2|\mathbf{x}) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2} \tag{2.15}$$

$$\text{cov}(b_1, b_2|\mathbf{x}) = \sigma^2 \left[\frac{-\bar{x}}{\sum (x_i - \bar{x})^2} \right] \tag{2.16}$$

For a simple linear regression model:

$$y = \beta_0 + \beta_1 x + \epsilon,$$

the design matrix X is:

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

Thus, $X'X$ becomes:

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

The inverse of $X'X$ is:

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

2. Variance-Covariance Matrix of b:

The covariance matrix for $b = [b_0, b_1]'$ is given by:

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

From this matrix, we can extract the variances and covariance of b_0 and b_1:

- Variance of b_0: The top-left element.
- Variance of b_1: The bottom-right element.
- Covariance between b_0 and b_1: Either off-diagonal element.

3. Variance of $b_1|X$:

The variance of $b_1|X$ corresponds to the bottom-right element of the covariance matrix:

$$\text{Var}(b_1|X) = \sigma^2 \frac{N}{N\sum x_i^2 - (\sum x_i)^2}.$$

Simplify using the relationship:

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

where $\bar{x} = \frac{\sum x_i}{N}$. Thus:

$$\text{Var}(b_1|X) = \sigma^2 \frac{1}{\sum (x_i - \bar{x})^2}.$$

This matches Equation (2.15).

4. Variance of $b_0|X$:

The variance of $b_0|X$ corresponds to the top-left element of the covariance matrix:

$$\text{Var}(b_0|X) = \sigma^2 \frac{\sum x_i^2}{N(\sum x_i^2 - (\bar{x})^2)}.$$

Simplify using similar steps as above, leading to:

$$\text{Var}(b_0|X) = \sigma^2 \left[\frac{\sum x_i^2}{N(\sum (x_i - \bar{x})^2)} \right].$$

This matches Equation (2.14).

5. Covariance Between $b_0|X$ and $b_1|X$:

The covariance between $b_0|X$ and $b_1|X$ corresponds to either off-diagonal element of the covariance matrix:

$$\text{Cov}(b_0, b_1|X) = -\sigma^2 \frac{\sum x_i}{N(\sum (x_i - \bar{x})^2)}.$$

Rewriting using the definition of $\bar{x} = \frac{\sum x_i}{N}$:

$$\text{Cov}(b_0, b_1|X) = \sigma^2 \left[\frac{-\bar{x}}{\sum (x_i - \bar{x})^2} \right].$$

HW0324Q3 (C05Q03)

5.3 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		Mean dependent var		6.19434
S.E. of regression		S.D. dependent var		6.39547
Sum squared resid	46221.62			

a.

i. The t-statistic for *b1*

- Formula: t = coefficient / standard error
- From the table, the coefficient (b1) is 1.4515 and the standard error is 2.2019.
- Calculation: t = 1.4515 / 2.2019 = 0.6592

ii. The standard error for *b2*

- Formula: Standard error = coefficient / t-statistic
- From the table, the coefficient (b2) is 2.7648, and the t-statistic is 5.7103.
- Calculation: Standard error = 2.7648 / 5.7103 = 0.4842

iii. The estimate *b3*

- From the table, the standard error is 0.3695, and the t-statistic is -3.9376.
- Formula: Coefficient = t-statistic * standard error

- Calculation: -3.9376 * 0.3695 = -1.455

iv. R^2

- Formula: R2 = 1 - (SSR/TSS). We know SSR (Sum of Squared Residuals) = 46221.62, but we need to calculate TSS (Total Sum of Squares).
- We know that TSS = (n-1) * (SD of dependent variable)^2
- So, TSS = (1200-1) * (6.39547)^2 = 48961.75
- Calculation: R2 = 1 - (46221.62/48961.75) = 0.05596 = 0.056

v. $\hat{\sigma}$ (Standard Error of the Regression)

- To find the S.E. of regression, we need to use the following formula: $\hat{\sigma}$ = sqrt(SSR / (n-k-1)), where SSR is the sum of squared residuals, n is the number of observations, and k is the number of independent variables.
- In this case, SSR = 46221.62, n = 1200, and k = 4 (ln(TOTEXP), NK, AGE, constant).
- $\hat{\sigma}$ = sqrt(46221.62 / (1200 - 4 - 1)) = sqrt(46221.62 / 1195) = sqrt(38.679) = 6.22

b.

From the result from a and the data in the table, b2 b3 and b4 are 2.7648, -1.455, and -0.1503. The interpretation are below:

B2: If total expenditure increases by 1%, the percentage of the household budget spent on alcohol is predicted to increase by approximately 2.7648 percentage points, holding other factors constant

B3: For each additional child in the household, the percentage of the household budget spent on alcohol is predicted to decrease by approximately 1.455 percentage points, holding other factors constant

B4: For each additional year of age of the household head, the percentage of the household budget spent on alcohol is predicted to decrease by approximately 0.1503 percentage points, holding other factors constant

c.

Confidence Interval = b ± (critical value * standard error)

Where:

- b is the estimated coefficient
- standard error is the standard error of the coefficient
- critical value is the t-value for the desired confidence level (95%) and degrees of freedom.

1. Gather the Information:

- *b4* (Coefficient for AGE): -0.1503
- Standard Error for *b4*: 0.0235
- Degrees of Freedom: n - k - 1 = 1200 - 4 - 1 = 1195

2. Find the Critical Value:

- Since the degrees of freedom (1195) are large, we can use the z-score for a 95% confidence level as an approximation of the t-value.
- For a 95% confidence level, the z-score is approximately 1.96.

3. Calculate the Confidence Interval:

- Lower Limit: -0.1503 - (1.96 * 0.0235) = -0.1503 - 0.04606 = -0.19636
- Upper Limit: -0.1503 + (1.96 * 0.0235) = -0.1503 + 0.04606 = -0.10424

Therefore, the 95% confidence interval for *β4* is approximately (-0.1964, -0.1042).

4. Interpretation:

We are 95% confident that the true value of *β4* (the effect of age on the percentage of the household budget spent on alcohol) lies between -0.1964 and -0.1042.

c.

The coefficients for ln(TOTEXP), NK, and AGE are all statistically significant at the 5% level because their p-values are less than 0.05. This means we have strong evidence to conclude that these variables have a statistically significant impact on the percentage of the household budget spent on alcohol (WALC).

The constant term (C) is not statistically significant at the 5% level because its p-value is greater than 0.05. This means we do not have strong evidence to conclude that the constant term is different from zero

d.

1. Define the Hypotheses:

- Null Hypothesis (H0): $\beta 3 = -2$ (The addition of an extra child decreases the mean budget share of alcohol by 2 percentage points)
- Alternative Hypothesis (H1): $\beta 3 \neq -2$ (The decrease is not equal to 2 percentage points)

2. Test Statistic:

We will use a t-test statistic to test this hypothesis. The formula is:

t = (b3 - hypothesized value) / standard error of b3

In this case:

- b3 (estimated coefficient for NK) = -1.455
- Hypothesized value = -2
- Standard error of b3 = 0.3695

Therefore:

t = (-1.455 - (-2)) / 0.3695 = (0.545) / 0.3695 = 1.475

3. Determine the Critical Value and Rejection Region:

- Significance Level (α) = 0.05
- Degrees of Freedom = n - k - 1 = 1200 - 4 - 1 = 1195
- Since this is a two-tailed test, we need to divide the significance level by 2 ($\alpha/2 = 0.025$).

- Using a t-table or calculator, we find the critical t-value for $\alpha/2 = 0.025$ and 1195 degrees of freedom. Since the degrees of freedom are large, we can use the z-score as an approximation. The critical z-value is approximately 1.96.
- Rejection Region: We reject the null hypothesis if the absolute value of our calculated t-statistic is greater than the critical t-value ($|t| > 1.96$).

4. Compare the Test Statistic to the Critical Value:

- Our calculated t-statistic is 1.475.
- The absolute value of the test statistic is $|1.475| = 1.475$.
- Since 1.475 is not greater than 1.96, we do not reject the null hypothesis.

5. Conclusion:

At a 5% significance level, we fail to reject the null hypothesis. There is not enough evidence to conclude that the addition of an extra child decreases the mean budget share of alcohol by an amount different from 2 percentage points. We cannot reject the claim that the decrease is 2 percentage points.

HW0324Q5 (C05Q23)

5.23 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

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

a.

beta 2 = negative

beta 3 = positive

beta 4 = positive

b.

```
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 signs of QUANT and TREND does not follow my expectation.

c.

R^2 = 0.5097

d.

Null Hypothesis (H0): There is no significant relationship between the quantity sold and the price per gram. Mathematically, this can be represented as:

H0: β2=0

This means that the coefficient for quant in the regression model is zero, indicating no effect of quantity on price.

Alternative Hypothesis (H1): There is a significant negative relationship between the quantity sold and the price per gram. Mathematically:

$$H1: \beta 2 < 0$$

This suggests that as the quantity increases, the price per gram decreases.

e.

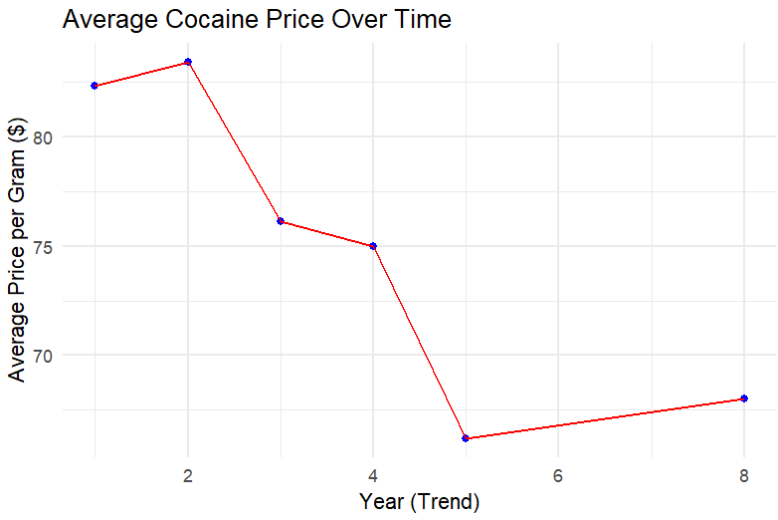
To test the hypothesis that the quality of cocaine has no influence on expected price against the alternative that a premium is paid for better-quality cocaine, we need to examine the coefficient for qual in the regression model.

Hypotheses:

- **Null Hypothesis (H_0):** $\beta 3 = 0$ (No relationship between quality and price).
- **Alternative Hypothesis (H_1):** $\beta 3 > 0$ (Positive relationship: higher quality leads to higher prices).

```
[1] "Coefficient for qual:  0.1162"
> print(paste("One-sided p-value for qual: ", qual_p_value_one_si
ed))
[1] "One-sided p-value for qual:  0.284996009969415"
>
> # Interpret the results
> if (qual_p_value_one_sided < 0.05 & qual_coefficient > 0) {
+   print("Reject H₀: There is a significant positive relationship
between quality and price.")
+ } else {
+   print("Fail to reject H₀: No significant positive relationship
between quality and price.")
+ }
[1] "Fail to reject H₀: No significant positive relationship betwe
en quality and price."
```

f.




```
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
```

In early year the quality of cocaine might be exceed the term of quantity and trend, while after 2nd year, the impact of trend to price will be significant. By the way, at 8th year, the quality of cocaine might improve significantly or there are external factors that affect the model leading to the price per gram increases.